**THÈSE DE DOCTORAT DE L'UNIVERSITE PARIS-SUD**

Mention : Informatique

Présentée par

**Tifanie BOUCHARA**

---

# Comparaison et combinaison de rendus visuels et sonores pour la conception d'interfaces homme-machine : des facteurs humains aux stratégies de présentation à base de distorsion.

---

Préparée au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (CNRS UPR 3251)
et soutenue le 29 octobre 2012 devant le jury composé de :

Rapporteurs :

| | |
|---|---|
| Stéphane Natkin | Professeur, CEDRIC, CNAM |
| Isabelle Viaud-Delmon | Directeur de Recherche au CNRS, IRCAM |

Examinateurs :

| | |
|---|---|
| Catherine Guastavino | Associate Professor, McGill University & CIRMMT |
| Laurence Nigay | Professeur, LIG, Université Joseph Fourier |
| Jean-Claude Martin | Professeur, LIMSI, Université Paris-Sud |

Directeurs :

| | |
|---|---|
| Christian Jacquemin | Professeur, LIMSI, Université Paris-Sud |
| Brian F.G. Katz | Chargé de Recherche au CNRS (HDR), LIMSI |

# Remerciements

J'ai longtemps cru que je n'arriverais pas au bout de cette thèse. Sans aucun doute, cela aurait été vrai si j'avais été seule. Mais j'ai été accompagnée, tout au long de cet exercice, par différentes personnes qu'il m'est difficile de remercier à leur juste valeur en seulement quelques mots.

Tout d'abord, rien de tout cela n'aurait pu voir le jour sans le soutien de mes directeurs Christian Jacquemin et Brian F.G. Katz. Je les remercie pour avoir cru en moi dès mon premier stage de Master 2 Image et Son et pour avoir accepté de reconduire l'expérience sur une plus longue durée à travers cette thèse. Merci à tous les deux pour les nombreux conseils scientifiques et nombreuses relectures qui ont sans aucun doute aussi profité aux industries du papier et des manufactures d'encre. Je remercie Christian pour sa rigueur scientifique qui me poussent sans cesse à aller chercher « LA » problématique, ainsi que pour m'avoir recommandée pour plusieurs projets Art-Sciences. Je remercie Brian tant pour nos échanges d'idées que pour son soutien sans relâche, même dans les moments les plus difficiles. Merci aussi pour m'avoir considérée comme une collègue par moments et m'avoir fait découvrir la recette des gnocchis et les aventures du Disque-monde.

Je remercie également Patrick Le Quéré, directeur du LIMSI, ainsi que Christophe d'Alessandro et Jean-Paul Sansonnet pour m'avoir respectivement accueillie au sein du laboratoire et intégrée aux équipes AA et AMI.

Un grand merci aux rapporteurs et aux examinateurs de cette thèse pour leur lecture attentive du manuscrit, leurs remarques qui m'ont permis de l'améliorer, et surtout pour avoir su enrichir encore ma réflexion, à travers les discussions qui ont suivi, en proposant de nouvelles pistes de recherche.

Je tiens à exprimer ma profonde gratitude envers Catherine Guastavino pour m'avoir invitée au Multimodal Interaction Lab. En plus d'y avoir découvert les domaines passionnants de la psychologie expérimentale et de la perception sonore et multisensorielle, c'est une aventure humaine incroyable que j'ai pu vivre là-bas. J'en profite pour remercier les collègues du MIL, notamment Amandine et Maryse, mais aussi Charles-Antoine, Ilja et Adrien, pour les excellents moments passés ensemble et pour m'avoir fait profiter un maximum de cette région du monde incroyable qu'est le Québec.

Merci aux collègues de l'équipe AA tant pour les séances musicales que pour les séances de brainstorming scientifique. En particulier, je remercie mes frères de thèse, Marc et Gaëtan, pour m'avoir supportée (ou suivie ;-) ?) depuis ATIAM et m'avoir fait découvrir leurs musiques, du rap à l'électro, à travers le couloir. Merci à mes collègues de bureau Nathalie et David pour leur nombreux conseils, tant méthodologiques (« Pourquoi ? Comment ? Jusqu'où ? Et après ? ») qu'informatiques, et pour les nombreuses discussions philosophiques, artistiques ou

pédagogiques, souvent les trois en même temps. Merci à Paul pour m'avoir amenée sur les toits de Paris, à Lionel pour sa main verte, à Sylvain pour ses conseils en Max/MSP, à Christophe et Albert pour leurs connaissances en traitement de la parole, à David pour ses gadgets Apple, à Mathieu pour sa pratique des pédales de loop, à Marc pour avoir de si bons goûts musicaux...

Merci aussi aux membres des équipes AMI et CPU qui, bien que plus éloignés géographiquement de quelques mètres, m'ont apporté de nombreux conseils. En particulier, merci à Matthieu pour son aide en programmation graphique, à Céline et Malika pour leur aide sur l'analyse des tests perceptifs, à Jonathan et Frédéric pour leur aide en interaction homme-machine, à Mathieu alias Udud pour son aide en LaTeX et OpenCV. Merci aussi à l'équipe technique, aux membres de l'équipe AMIC et à Laurent, pour leur aide en programmation web, en Python et pour tout le reste.

Je remercie aussi mes collaborateurs pour m'avoir tant apporté. Merci à Ilja pour ses connaissances en Matlab, à Bruno pour sa base de données de sons d'environnement, à Marina pour les vidéos d'interview du forum juridique, à Eric Bilinski pour les analyses automatiques. Merci à l'équipe technique du projet ORA (Orgue et Réalité Augmentée) et au groupe GAVIP (Gestural Auditory and Visual Interactive Platform) pour les expériences Art-Sciences extérieures à la thèse qui m'ont aussi permis de souffler de temps en temps.

Merci à tous les participants, en particulier un énorme merci à tous mes amis et membres de la famille qui ont testé les expériences pilotes en avant-première, parfois même à en pleurer (désolée maman).

Merci aussi à toutes les personnes qui ont relu mon manuscrit pour en ôter un maximum de fautes d'orthographe, à savoir ma grand-mère, Aurélie, Béatrice, Diane et Guillaume.

Merci à Anne-Lyse, ma partenaire du Bureau des Doctorants, ainsi que tous les doctorants et collègues du LIMSI pour avoir rendu ces années au labo si agréables.

Un énorme MERCI à mes amis, Amandine, Aurélie, Béatrice, Gaëtan, Guillaume, Lise, Mathieu, Marc, et Marina et à tous les autres dont le soutien sans faille m'a permis d'aller jusqu'au bout. Un merci tout spécial à Rémi pour m'avoir redonnée confiance en moi et pour avoir été si compréhensif durant cette dernière année.

Enfin, merci à ma famille, mes parents, mon frère et tous les autres, pour avoir cru en moi, m'avoir soutenue, m'avoir amenée jusque là.

# Table des matières

# Table des figures

# Liste des tableaux

# Nomenclatures et abréviations

| | |
|---|---|
| $\boldsymbol{\alpha_{int}}$ | Azimut des sources sur le périmètre interne de la lentille |
| $\boldsymbol{\alpha_{max}}$ | Azimut des sources sur le périmètre externe de la lentille |
| **A** | Condition Audio seul |
| **ANOVA** | Analyse de variance (pour *ANalysis Of VAriances*) |
| **AV** | Condition Audiovisuelle, flou appliqué sur les deux composantes à la fois |
| $\mathbf{AV^{neut}}$ | Condition Audiovisuelle, flou appliqué uniquement à l'audio (visuel neutre) |
| $\mathbf{A^{neut}V}$ | Condition Audiovisuelle, flou appliqué uniquement au visuel (audio neutre) |
| **AVc** | Condition Audiovisuelle avec une image congruente |
| **AVi** | Condition Audiovisuelle avec une image incongruente |
| **AVn** | Condition Audiovisuelle avec une image neutre |
| **AV-FL** | Présentation Audiovisuelle avec la lentille *Fisheye* |
| **AV-PZ** | Présentation Audiovisuelle avec la méthode Pan&Zoom |
| **B+T** | Technique Bifocal+Transparence |
| **c** | Constante empirique pour le calcul de la distorsion de volume sonore |
| **DTW** | Algorithme d'alignement temporel (pour *Dynamic Time Warping*) |
| $\mathbf{f_c}$ | Fréquence de coupure d'un filtre passe-bas, réglage du niveau de flou audio |
| **FL** | Technique de lentille en œil-de-poisson (pour *Fisheye Lens*) |
| **HNR** | Rapport d'harmonicité (pour *Harmonics-to-Noise Ratio*) |
| **HRTF** | Fonction de transfert relative à la tête (pour *Head Related Transfer Function*) |
| **IHM** | Interface ou Interaction Homme-Machine |
| **ILD** | Différences interaurales d'intensité (pour *Interaural Level Difference*) |
| **ITD** | Différences interaurales de temps (pour *Interaural Time Difference*) |
| **MFCC** | Coefficients cepstraux (*Mel Frequency Cepstral Coefficients*) |
| **MLE** | Maximum de vraisemblance (pour *Maximum Likelihood Estimation*) |
| **NPR** | Rendus Non Photoréalistes (pour *Non Photorealistic Rendering*) |
| **RSB** | Rapport Signal-sur-Bruit (SNR en anglais pour *Signal-to-Noise Ratio*) |
| **RT** | Temps de réponse (pour *Response Time*) |
| $\mathbf{p_{Assoc}}$ | Taux d'association son-image |
| $\mathbf{p_{vivant}}$ | Probabilité qu'un son soit identifié comme « *vivant* » |
| **PZ** | Technique de Pan&Zoom |
| **r** | Distance au centre de la lentille dans la fenêtre de rendu |
| **r1** | Distance au centre de la lentille à l'échelle de la texture (échelle 1) |
| $\mathbf{rad_{int}}$ | Rayon interne de la lentille grossissante en pixels |
| $\mathbf{rad_{ext}}$ | Rayon externe de la lentille grossissante en pixels |
| **RM-ANOVA** | Analyse de variance à mesures répétées (pour *Repeated Measures ANalysis Of VAriances*) |
| **RSE** | Effet de cible redondante (pour *Redundant Signal Effect*) |
| **RSVP** | *Rapid Serial Visual Presentation* |
| **SDOF** | Technique de profondeur de champ sémantique (pour *Semantic Depth-Of-Field*) |
| **V** | Condition Visuel seul |
| $\mathbf{v_{min}}$ | Volume sonore minimal, appliqué aux sources à l'extérieur de la lentille |
| **VBAP** | Technique de spatialisation sonore *Virtual Based Amplitude Panning* |
| **V-FL** | Présentation Visuelle avec la lentille *Fisheye* |
| **V-PZ** | Présentation Visuelle avec la méthode Pan&Zoom |
| **WFS** | Technique de spatialisation sonore *Wave Field Synthesis* |
| **ZUI** | Interfaces zoomables ou multi-échelles (pour *Zoomable User Interface*) |
| **ZR** | Facteur de zoom (pour *Zoom Ratio*) |
| **%Correct** | Taux de réponses correctes |
| $\boldsymbol{\sigma}$ | Rayon de flou gaussien, paramètre de réglage du niveau de flou visuel |

Chapitre 1

# Introduction

*The challenge is about taking things that are infinitely complex and making them simpler and more understandable.*
Robert Greenberg, 2006

**Sommaire**



## 1.1 L'exploration de grandes collections de documents multimédia

Depuis l'arrivée des *smartphones*, appareils photographiques, caméras et autres appareils de capture numérique grand public, les bases de données personnelles ne cessent d'augmenter. Cette grande quantité d'information est cependant rarement organisée autrement qu'en dossiers/sous-dossiers et le nom des fichiers se limite souvent à une suite de chiffres attribuée automatiquement par l'appareil de capture (comme le nom « P134005.JPG » pour une photographie). Ce manque d'organisation et d'étiquetage rend laborieuse la recherche d'un document particulier dans la collection de documents disponibles et oblige souvent l'utilisateur à survoler la collection en affichant les documents d'un dossier un par un. La mise en place d'outils de présentation adaptés est nécessaire pour permettre à l'utilisateur de parcourir plus rapidement sa collection personnelle. On pourrait penser que le problème est différent pour des bases de données stockées sur Internet, accessibles par des réseaux de partage comme *Youtube* ou *DailyMotion*, puisque les données partagées par les internautes sont alors étiquetées et analysées de sorte qu'il est possible de faire une recherche par mot-clé ou par similarité. Ces méthodes de recherche où l'utilisateur formule une requête sont étudiées dans le domaine de la recherche d'information (*Information Retrieval*) mais ne nous concernent pas ici car elles présentent des limites qui conduisent au même problème de présentation que les collections non organisées. En effet, d'une part, l'utilisateur peut ne pas connaître ce qu'il cherche réellement ou ne pas pouvoir formuler facilement une requête adaptée, et d'autre part, les résultats obtenus par un moteur de

recherche sont souvent nombreux[1]. Dans un cas comme dans l'autre, il faudra alors présenter les différents documents possibles, qu'ils aient été sélectionnés par le moteur de recherche ou non, de façon la plus adaptée possible pour que l'utilisateur puisse parcourir cette collection ou sous-collection rapidement.

C'est sur cette problématique de la présentation et de l'accès à de nombreux documents numériques (visuels, sonores ou multimédia) que nous avons consacré notre travail de thèse.

Ce projet s'inscrit dans le domaine de l'interaction homme-machine (IHM) et se retrouve donc à l'intersection de deux disciplines scientifiques : l'informatique et la psychologie expérimentale. En effet, le problème de la présentation de documents fait référence à un sous domaine de l'IHM, à savoir l'interaction en sortie, qui étudie les mécanismes qui permettent à la machine de communiquer des informations à l'utilisateur. Cette interaction peut être vue comme un lien de l'ordinateur vers l'utilisateur, et nous étudierons chaque extrémité du lien. Ainsi l'affichage des informations par l'ordinateur, soit par un rendu graphique soit par un rendu sonore, évoque des aspects informatiques, tandis que la façon dont l'utilisateur perçoit ce rendu réfère à la psychologie expérimentale. Notre objectif est d'adapter la présentation des données en fonction de la perception de l'utilisateur pour optimiser l'interaction.

## 1.2 Démarche scientifique

Pour répondre aux problèmes d'affichage de multiples documents visuels, de nombreuses stratégies de présentation ont été proposées dans les interfaces visuelles. Elles permettent à l'utilisateur d'explorer rapidement une collection et d'accéder à un maximum de données en un minimum de temps. Ces interfaces reposent sur une présentation simultanée de plusieurs documents et sur des distorsions et des changements d'échelle pour mettre en relief l'information la plus pertinente. En revanche, alors même que de nombreux documents numériques ont un contenu sonore (vidéos, fichiers .mp3, etc...), peu de stratégies de présentation permettent d'accéder rapidement à ce contenu.

Notre objectif est de fournir des stratégies de présentation audio et audiovisuelles adaptées pour améliorer la conception des interfaces d'exploration de grandes collections multimédia.

Pour cela, nous avons orienté notre recherche selon trois axes. Tout d'abord, il s'agit de définir des équivalents auditifs aux stratégies visuelles disponibles, de sorte que l'utilisateur puisse accéder à plusieurs documents sonores simultanément. Ensuite, il s'agit d'étendre ces stratégies à l'audiovisuel pour la présentation de documents multimédia. Le passage à l'audiovisuel soulève une nouvelle question sur la manière d'optimiser la combinaison des modalités audio et visuelles ainsi que la combinaison des stratégies proposées dans chaque modalité. Enfin, il faut encore proposer des interfaces d'exploration de bases de données, fondées sur ces stratégies, qui soient adaptées à l'utilisateur et à ses capacités. Afin d'éviter la surcharge perceptive et cognitive induite par la présentation de trop nombreux documents en même temps, nous orientons nos recherches sur les processus perceptifs et attentionnels impliqués dans l'observation d'objets audiovisuels concurrents, en insistant sur les interactions entre les deux modalités.

1. Le nombre de résultats apportés par le moteur de recherche correspond, d'une part, aux résultats qui répondent totalement à la requête, mais aussi à ceux qui y répondent de façon plus éloignée afin de pallier la mauvaise formulation éventuelle de la requête.

## 1.3 Organisation du manuscrit

Le manuscrit est organisé en trois parties. La première partie (chapitres 2 et 3) donne le cadre général de l'étude ainsi que nos motivations. Il s'agit d'une analyse des travaux pré-éxistants sur les stratégies de présentation déjà disponibles et sur la perception et l'attention humaine. Le reste du document se concentre sur les apports de notre travail. La deuxième partie du manuscrit (chapitre 4) est orientée sur le développement de stratégies de présentation audiovisuelles en s'inspirant des stratégies audio et visuelles déjà existantes. La troisième partie du manuscrit (chapitres 5 et 6) explore plus en profondeur les phénomènes attentionnels et perceptifs humains et propose de nouvelles stratégies de présentation audiovisuelles qui exploitent ces phénomènes. Le détail de chacune des parties est donné dans les sections qui suivent.

### 1.3.1 Présentation du contexte

Le chapitre 2 analyse les travaux sur le rendu des interfaces d'accès à l'information dans de grandes collections audiovisuelles. Il expose les différentes stratégies de présentation visuelles, auditives ou dédiées au multimédia, en se focalisant principalement sur celles où les documents sont présentés simultanément. Nous démontrerons que les outils actuels sont peu adaptés à la présentation de documents audiovisuels tels que des vidéos musicales. Ce chapitre nous donnera également l'occasion de justifier en quoi une présentation audiovisuelle peut être une solution plus adaptée qu'une présentation unimodale. Enfin les questions sur l'utilisation de l'audio dans une interface d'exploration de collections audiovisuelles seront introduites.

L'aspect utilisateur sera abordé au chapitre 3. Nous y présenterons les phénomènes perceptifs et attentionnels mis en jeu lors d'une tâche de recherche, visuelle ou auditive, où l'utilisateur doit retrouver un document particulier présenté conjointement à plusieurs autres documents audio ou visuels. Certains paramètres augmentent la saillance d'un objet qui attire alors, sans effort particulier, l'attention de l'utilisateur. L'analyse bibliographique mettra en relation ces paramètres et leur exploitation dans les stratégies de présentation présentées au chapitre 2. Finalement, ce chapitre abordera aussi les études sur la perception multisensorielle, afin d'examiner l'intérêt d'une présentation multimodale plutôt qu'unimodale.

### 1.3.2 Premières stratégies de présentations audiovisuelles

Reprenant les stratégies audio et visuelles du chapitre 2, nous avons établi un modèle théorique de rendu audiovisuel que nous présenterons au chapitre 4. Deux méthodes, appelées Pan&Zoom et Lentille Fisheye, ont été implémentées en combinant des méthodes unimodales. La première méthode utilise un rendu sans distorsion visuelle, tandis que la seconde méthode utilise un rendu audio et un rendu visuel distordus. Une étude d'utilisabilité sur une application de recherche de vidéo a été menée afin d'évaluer l'apport de chaque modalité, et en particulier de l'audio, dans la présentation audiovisuelle obtenue. De plus, en comparant les performances et les jugements subjectifs des utilisateurs sur ces deux méthodes, nous proposons des règles pour optimiser la combinaison audiovisuelle. Cette étude est publiée dans [Bouchara *et al.* 2010b] :

- Tifanie Bouchara, Catherine Guastavino, Brian F.G. Katz, Christian Jacquemin. « Audio-Visual Rendering for Multimedia Navigation », Proceedings of the 16th International Conference on Auditory Display (ICAD'2010), Washington D.C., États-Unis, juin 2010.

### 1.3.3 Etude et exploitation de facteurs perceptifs et attentionnels

Le chapitre 5 est consacré à une étude purement perceptive sur l'identification des sons d'environnement dans un milieu bruité, en présence ou non d'un contexte visuel. Le bruit simule ici la présence de plusieurs sources simultanées telles qu'on peut les retrouver dans une interface où les documents audio et audiovisuels sont présentés simultanément. Au-delà des buts premiers de la thèse, l'étude évalue l'influence de la congruence sémantique entre les composantes visuelle et sonore pour la reconnaissance d'objets et s'attache à approfondir les connaissances sur la perception auditive des sons d'environnement. Cette étude est décrite dans la publication [Bouchara *et al.* 2010a] :

- Tifanie Bouchara, Bruno L. Giordano, Ilja Frissen, Brian F.G. Katz, Catherine Guastavino. « Effect of Signal-to-Noise Ratio and Visual Context on Environmental Sound Identification », Proceedings of the $128^{th}$ convention of the Audio Engineering Society (AES $128^{th}$), Londres, Grande-Bretagne, mai 2010.

Le chapitre 6 revient sur la notion de saillance et sur les processus attentionnels impliqués dans la recherche d'un objet d'intérêt parmi plusieurs objets non pertinents. Reprenant les observations selon lesquelles un objet visuel net attire l'attention au milieu d'objets flous, nous avons étendu la notion de flou visuel aux modalités auditives et audiovisuelles par analogie. L'étude présentée dans ce chapitre est organisée autour d'une série d'expériences perceptives, pour évaluer l'effet d'attraction dans chaque modalité en fonction du niveau de flou, ainsi que l'effet d'attraction obtenue dans une présentation audiovisuelle. Les résultats de cette étude ont conduit aux publications suivantes [Bouchara *et al.* 2012; Bouchara et Katz 2012] :

- Tifanie Bouchara, Christian Jacquemin, Brian F.G. Katz. « *Cueing Multimedia Search with Audio-Visual Blur* ». En phase de première révision pour ACM Transactions on Applied Perception. 2012.
- Tifanie Bouchara, Brian F.G. Katz, Christian Jacquemin. « Guidage attentionnel à base de flou audiovisuel pour la conception d'interfaces multimodales », Conférence sur l'Ergonomie et l'Informatique Avancée (Ergo'IHM 2012), Biarritz, France, octobre 2012. Prix de la meilleure communication scientifique.
- Tifanie Bouchara, Brian F.G. Katz. « Redundancy gains in audio-visual search ». Proceedings of the International Multisensory Research Forum (IMRF 2012). Seeing and Perceiving, vol.25 (S1), p.181, 2012.

Finalement, le chapitre 7 est dédié à la conclusion. Ce chapitre résume les contributions de la thèse et reprend les résultats obtenus sur les différentes études des chapitres 4, 5 et 6. Des suggestions de travaux futurs y sont également présentées.

## 1.4 Contributions de la thèse

Pour résumer, nos principales contributions sont d'avoir développé plusieurs stratégies de présentation audiovisuelles :

- nous avons modélisé et implémenté deux stratégies audiovisuelles multi-échelles, Pan&Zoom et lentille grossissante, obtenues par modification des paramètres de taille visuelle et de volume sonore ;
- nous avons défini des effets de flou audio et de flou audiovisuel, obtenus par filtrage fréquentiel, et montré qu'ils permettent d'augmenter la saillance d'un objet net alors mis en avant parmi plusieurs objets flous, et ainsi d'attirer vers lui l'attention de l'utilisateur de façon involontaire donc sans effort.

Ces stratégies ont également servi de sujets d'étude sur la multimodalité et la perception multisensorielle, ce qui nous a permis de :

- montrer l'apport de la modalité auditive, ajoutée à la modalité visuelle, dans une présentation de plusieurs documents multimédia simultanés, ainsi que l'avantage d'une présentation multimodale lorsque l'audio est dégradé ;
- prouver qu'il est possible de diriger l'attention visuelle, l'attention sonore et l'attention audiovisuelle d'un utilisateur pour lui faciliter une tâche de recherche ;
- découvrir que combiner les modalités auditive et visuelle n'implique pas d'appliquer les mêmes traitements dans chacune des modalités.

## 1.5 Exemples sonores et vidéos

Nous proposons, tout au long de ce manuscrit, différents exemples audio et multimédia. Ils sont repérés dans la marge par le symbole  1.0 pour un exemple audio (ici l'exemple audio 1.0) et par le symbole  1.0 pour un exemple vidéo (ici l'exemple vidéo 1.0). Cela permet d'illustrer par du son des concepts liés à la modalité auditive, concepts souvent difficiles à expliquer avec les illustrations visuelles courantes (photographies, graphiques, schémas, etc.). Nous invitons le lecteur à se référer à ces exemples disponibles à l'adresse Internet :
http://groupeaa.limsi.fr/media/tifanie/EXEMPLES_THESE/index.html.

# Première partie

# Contexte

Chapitre 2

# Stratégies de présentation dans les interfaces visuelles, audio et audiovisuelles

*There is no such thing as information overload. There is only bad design.*
Edward Tufte

## Sommaire



Les nouvelles technologies permettent de créer, stocker et échanger de plus en plus de documents numériques. Par conséquent, nos disques durs contiennent des collections de photographies, de vidéos ou de fichiers audio, de plus en plus larges. Afficher tous ces documents simultanément sur un même écran, avec une taille suffisante, n'est alors pas possible. Ce manque de place, qualifié de problème d'espace écran, a donné lieu à de nombreuses études pour permettre à l'utilisateur d'accéder aux données et d'explorer ces grandes collections de documents. Nous présentons ici quelques unes de ces stratégies de présentation déjà disponibles pour l'exploration de collections.

Ce chapitre s'organise autour de trois axes en considérant séparément trois modalités de présentation : visuelle, auditive et audiovisuelle. Nous expliquerons dans un premier temps les

enjeux de la visualisation d'informations et les différentes techniques dédiées à l'exploration de documents visuels (section 2.1). Ensuite les sections 2.2 et 2.3 présenteront les stratégies de présentation de documents respectivement sonores et multimédia.

## 2.1 Stratégies de présentation en visualisation d'information

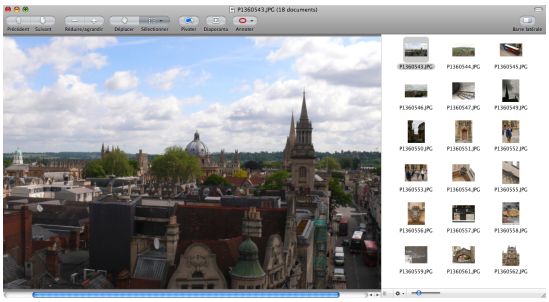
L'exploration de collections de documents visuels repose sur un ensemble de stratégies de présentation dont le but est de rendre disponible un maximum d'informations en un minimum de temps. De ce fait, la plupart des stratégies visuelles optent pour une présentation simultanée de nombreux documents. Le problème d'espace écran mentionné auparavant limite soit la taille des documents, soit le nombre de documents que l'on peut afficher simultanément. Plusieurs solutions ont alors été proposées dans le domaine de la visualisation d'informations. Ces différentes techniques sont basées sur une représentation à plusieurs niveaux de détail des documents. On parle alors d'***interfaces zoomables*** (ou **ZUI** pour Zoomable User Interfaces) ou interfaces multi-échelles [Cockburn *et al.* 2008; Spence 2001]. Certaines de ces méthodes agissent sur l'espace de représentation soit en n'affichant qu'une partie des ou du document (comme dans une présentation page par page) soit en déformant cet espace (*lentille grossissante*). D'autres stratégies utilisent une modification des propriétés visuelles de certains éléments pour mettre les éléments les plus intéressants en valeur (*Semantic Depth of Field*, *lentilles magiques*). Ces stratégies sont notamment utilisées pour l'exploration de collection photographiques [Plaisant *et al.* 1995; Christmann 2008]. Elles ne sont toutefois pas réservées à l'affichage de documents numériques et se retrouvent dans d'autres contextes comme la représentation de données sur des graphiques, par exemple des arbres hiérarchiques avec la technique de *browser hyperbolique* [Lamping *et al.* 1995] ou l'outil *ChronoLens* pour l'exploration de graphiques temporels [Zhao *et al.* 2011]. Enfin ces techniques peuvent parfois servir à des fins artistiques, comme dans les *techniques de rendu expressif* [Kyprianidis *et al.* 2012].

### 2.1.1 Overview+Detail

Les ZUI à base d'***Overview+Detail*** présentent simultanément sur l'écran une vue large (overview) et une vue détaillée (vue principale) d'un même espace d'information. Les deux vues sont affichées dans deux fenêtres distinctes mises côte à côte. Dans l'exemple présenté figure 2.1a, l'application *Google Maps*[1] présente un plan de la ville de Paris avec une vue détaillée dans la plus grande fenêtre tandis qu'une autre fenêtre, en bas à droite, indique de quelle zone il s'agit. On utilise parfois le terme « vue en radar » pour parler de ce type de rendu. On retrouve les vues Overview+Detail aussi dans les applications d'exploration de collections : un document est affiché en grand dans l'espace central tandis que des miniatures (vignettes ou *thumbnail*) des autres documents sont affichées sur les bords (figure 2.1b). En général, l'overview est présentée sur un bord ou un coin de l'écran et peut être désactivée, ce qui laisse à penser qu'elle est moins utile que la vue détaillée. De plus, fusionner des informations séparées spatialement a un coût cognitif pour l'utilisateur qui aura donc tendance à se référer au détail et à l'overview alternativement.

1. L'application *Google Maps* est disponible en français à l'adresse https://maps.google.fr/.

**(a)** Juxtaposition d'un morceau de plan de Paris et d'une vue plus globale de la ville dans la fenêtre en bas à droite avec l'application Google Maps.



**(b)** Exploration d'une collection de photographies.

**Figure 2.1.** Deux exemples de présentation Overview+Detail.

### 2.1.2 Pan&Zoom

Deux types de zoom sont à différencier : le zoom géométrique et le zoom sémantique.

Le ***zoom géométrique*** permet à l'utilisateur de changer l'échelle, donc la taille d'une zone spécifique, appelée ***focus***, tandis que l'information en dehors de cette zone n'est pas affichée. Le mouvement de ***pan*** permet alors à l'utilisateur de repositionner la fenêtre sur une autre zone par translation. Dans une telle approche, le rendu est homogène (sans distorsion), mais il n'y a pas de vision globale. Cockburn et ses collègues [Cockburn *et al.* 2008] considèrent que la technique Pan&Zoom est une séparation entre le focus et le contexte, non plus spatiale comme pour la technique ***Overview+Detail***, mais temporelle puisqu'on alterne les deux types de vue. Son principal défaut apparaît dès qu'on l'utilise pour une vue détaillée : le contexte disparaît vite et une petite partie de l'information seulement est accessible. Ainsi, comme les utilisateurs ne peuvent pas voir la relation entre la partie visible et le reste de la structure du document ou de la collection, ils peuvent être désorientés par le manque de contexte visuel. On parle alors de brouillard ou ***Desert Fog***. De ce fait, dans certains cas, l'interaction peut nécessiter des mouvements supplémentaires : l'utilisateur doit faire un zoom arrière (*zoom out*) pour obtenir une vue d'ensemble avant de déplacer la zone de focus et ensuite re-zoomer (zoom avant ou *zoom in*) pour retrouver le niveau de détail souhaité ailleurs.

Les « ***space-scale diagrams*** » ont été définis par Furnas et Bederson [Furnas et Bederson 1995] pour expliciter les concepts de Pan&Zoom et offrir une visualisation des processus utiles dans les ZUI. Ce type de visualisation est présenté en figure 2.2 : la fenêtre de vue (a) peut être déplacée sur un axe vertical de bas en haut pour faire un zoom avant (et *vice versa* pour le zoom arrière) et sur un plan horizontal pour effectuer un mouvement de *pan*. Ainsi la vue (c) est une vue large du document tandis que la vue (b) est une vue zoomée sur une zone de l'espace et que la vue (d) présente une autre partie du document un peu moins agrandie que dans la vue (b).

Au contraire du zoom géométrique, le ***Zoom sémantique*** ne joue pas sur la taille des éléments mais sur leur représentation. A chaque niveau de détail correspond une nouvelle représentation plus ou moins détaillée. Cette technique réorganise également les informations au sein d'une

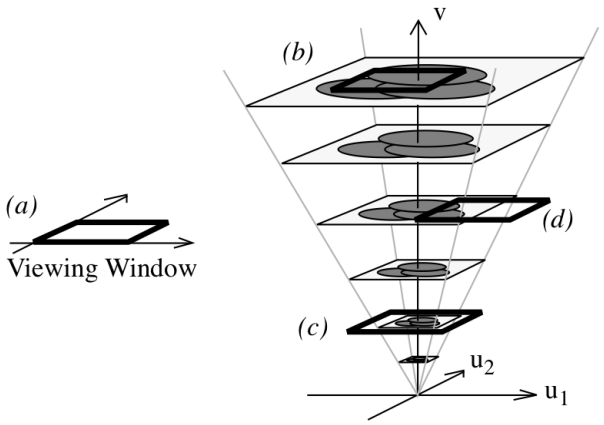



**Figure 2.2.** Représentation du déplacement de la fenêtre de vue dans un « *space-scale diagram* ». Extrait de [Furnas et Bederson 1995].

structure pour présenter à l'utilisateur uniquement l'information pertinente du niveau de détail choisi.

La ***présentation page par page*** peut être vue comme une méthode Pan&Zoom dans le sens où chaque page prend tour à tour la place de focus. Cette solution est surtout envisageable dans les cas où le contexte n'est que rarement utile, ou pour présenter des groupes de documents (hiérarchisation en dossier/sous-dossier par exemple). Cependant, même en ne présentant alors qu'un sous-ensemble de documents, une réflexion sur l'affichage des documents et sur les techniques d'interaction au sein d'une même page est nécessaire.

### 2.1.3 Techniques Focus+Contexte

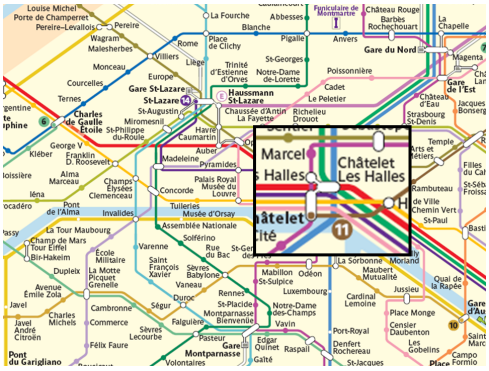
Au contraire des techniques de Pan&Zoom ou d'Overview+Detail, les techniques ***Focus+Contexte*** permettent d'afficher simultanément et dans le même espace la zone d'intérêt principal (focus) et l'environnement autour (contexte). Ce type de techniques est aussi appelée ***detail-in-context*** car la zone de détail est insérée directement dans le contexte. Il y a donc superposition des deux vues.

#### 2.1.3.1 Lentille grossissante bifocale

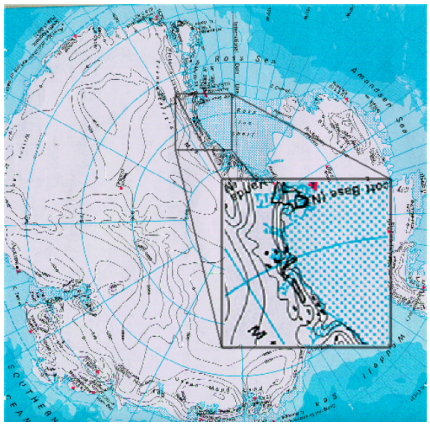
Une solution pour afficher le focus et le contexte simultanément et dans la même fenêtre consiste à superposer le focus au-dessus du contexte. On peut parler de ***lentille bifocale***. Les deux zones sont alors présentées simultanément sans distorsion et la position du focus indique le lien entre cette zone détaillée et la vue d'ensemble. Cependant cette méthode présente un défaut majeur car le focus masque alors une partie du contexte : une partie de l'information n'est alors pas disponible.

Pour éviter cette gêne, le ***DragMag Image Magnifier*** permet de déporter la vue du focus grâce à une lentille *DragMag* [Ware et Lewis 1995], aussi nommée « lentille Manhattan » [Carpendale et Montagnese 2001], dont un exemple de rendu est présenté figure 2.3b. Toutefois

le lien entre le focus et le contexte est alors plus difficile à faire à cause de la séparation spatiale entre les deux vues.




**(a)** Application d'une lentille bifocale à un plan de métro de Paris.



**(b)** Une lentille *DragMag* sur une carte de l'Antartique. Extrait de [Carpendale 2008].

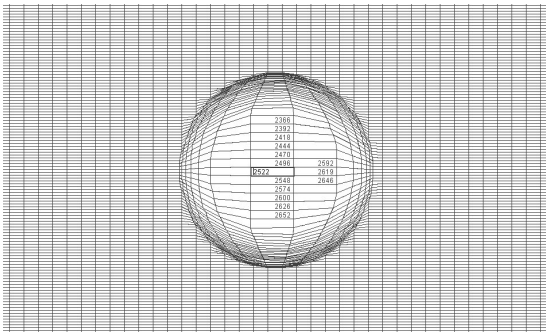
**Figure 2.3.** Deux exemples de rendu bifocal homogène, avec ou sans déport du focus.

#### 2.1.3.2 Lentille en œil-de-poisson, Fisheye View

Une autre option est de déformer l'espace de représentation. Leung et Apperley [Leung et Apperley 1994] fournissent une analyse de la littérature assez complète sur ces techniques à base de distorsions. Parmi ces techniques, la plus connue est la vue dite en ***œil-de-poisson*** (*Fisheye Views*). A l'origine [Furnas 1981], cette technique reposait sur la suppression des parties de document dont l'intérêt pour l'utilisateur ne dépassait pas un seuil fixé par l'utilisateur lui-même. On parlait alors d'affichage par filtrage. Le ***niveau d'intérêt*** de chaque objet ou élément (*DOI* pour *Degree of Interest*) a d'abord été conçu pour des informations hiérarchisées. Le niveau d'intérêt est alors relié à l'échelon de l'objet dans la hiérarchie. Le ***niveau de détail*** avec lequel un élément doit être affiché (*LOD* pour *Level of Detail*) dépend alors du niveau d'intérêt de l'objet et de la distance entre cet objet et la zone de focus.

Le concept de lentille en œil-de-poisson est généralisable [Furnas 1986] et Sarkar et Brown [Sarkar et Brown 1994] l'ont étendu à la distorsion graphique en 2D. Cette technique n'est alors plus une technique par filtrage mais une technique par déformation. La zone graphique de focus est élargie tandis que le reste de l'image est réduit proportionnellement à la distance euclidienne au centre de la lentille. Cette méthode permet d'obtenir une zone de focus avec un niveau de détail suffisant (vue zoomée ou élargie), tout en préservant la visualisation simultanée des zones de focus et de contexte, comme on peut le voir sur les exemples de la figure 2.4. Keahey et Robertson [Keahey et Robertson 1996] parlent de grossissement non-linéaire (*non linear magnification*) Parfois le grossissement est trop faible pour que la zone aggrandie soit facilement perçue. Il est alors possible d'indiquer la zone de grossissement par des informations supplémentaires, comme de l'ombrage, ce qui augmente l'impression de volume induit par le grossissement [Pietriga et Appert 2008].


**(a)** Application d'une lentille fisheye à un plan de métro de Paris.



**(b)** Utilisation d'une lentille fisheye sur un tableur de grande taille dans l'application FiCell (pour *Fisheye Cell*). Extrait de http://vernier.frederic.free.fr/FiCell/.

**Figure 2.4.** Deux exemples de rendu avec lentille fisheye.

#### 2.1.3.3 Superposition par transparence

Lecolinet et Pook [Lecolinet et Pook 2002] ont suggéré d'utiliser une ***technique par transparence*** dans laquelle le contexte est affiché sur une couche transparente au dessus du focus. A l'inverse, Lieberman affiche dans son ***Macroscope*** [Lieberman 1997] d'abord le contexte puis au-dessus le focus en transparence. Dans les deux cas la transparence permet de superposer le focus et le contexte simultanément sans masquage et sans distorsion. Cependant, du fait que la quantité d'information présentée est augmentée et que les deux vues sont difficiles à séparer l'une de l'autre, l'effort cognitif demandé à l'utilisateur est plus grand. Un exemple d'une telle superposition est présenté figure 2.5 pour l'exploration d'une carte géographique.

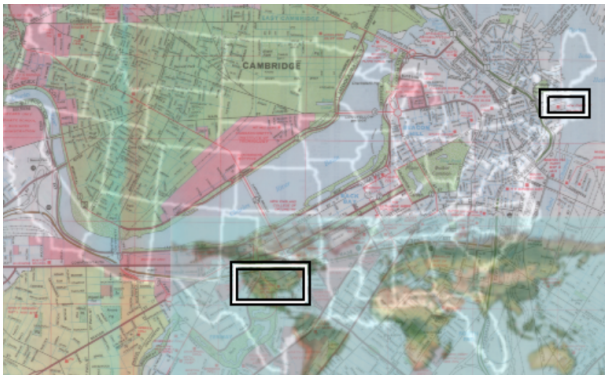



**Figure 2.5.** Exemple d'une présentation par transparence : trois couches de niveau de zoom différent sont superposées pour indiquer la position d'une carte routière de Cambridge sur une carte plus large des Etats-Unis et une carte mondiale. Extrait de [Lieberman 1997].

#### 2.1.3.4 Profondeur de champ sémantique, Semantic Depth-of-Field

La technique de ***profondeur de champ sémantique*** ou SDOF pour *Semantic Depth-of-Field* a été présentée par Kosara [Kosara 2001; Schrammel *et al.* 2003]. La SDOF s'appuie sur un effet perceptif : le regard est naturellement attiré par une partie d'image nette lorsque le reste de l'image est flou. Cet effet, largement utilisé en photographie avec la profondeur de champ pour mettre en relief les éléments d'une scène, peut donc facilement être appliqué ici en rendant flou le contexte ou la partie la moins pertinente pour l'utilisateur. Les détails n'apparaîtront plus que dans le focus qui attirera alors l'attention.

La figure 2.6 présente des applications de la SDOF à l'apprentissage du jeu d'échec (les pièces dont le joueur débutant doit tenir compte sont nettes) ou à de la sélection de texte.


**(a)** Apprentissage des règles aux échecs. **(b)** Mise en valeur de texte.

**Figure 2.6.** Exemples de présentation à base de profondeur de champ sémantique. D'après [Kosara 2001].

Comme pour les techniques par transparence, il n'y a ici ni déformation spatiale ni masquage. La séparation entre les deux zones dépendra du niveau de flou : plus le contexte est flou et plus le focus se démarque [Schrammel *et al.* 2003]. La démarquation entre le contexte et le flou peut se faire de manière progressive avec un niveau de flou paramétrable : plus l'information est floue moins elle est pertinente. Cela permet d'obtenir une hiérarchie étendue entre les éléments, et non plus seulement un effet binaire focus *versus* contexte. De plus, Rosenbaum et ses collègues [Rosenbaum et Schumann 2009] ont proposé d'étendre cette technique en ajoutant, au flou, de la désaturation (passage en échelle de gris sur la région de contexte). La distinction entre focus et contexte est alors renforcée et la zone de focus, nette et en couleur, attire d'autant plus l'attention.

### 2.1.4 Présentation séquentielle

Une autre façon de gagner du temps pour la recherche de documents visuels est d'afficher les documents un par un mais très rapidement. C'est la technique de ***Rapid Serial Visual Presentation*** ou RSVP [de Bruijn et Spence 2000; Spence 2002].

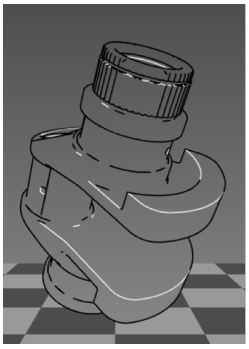
Le *« Coup d'œil » d'Apple*, appliqué à des photographies (figure 2.7), est un exemple de ce mode de rendu RSVP. Appuyer sur la flèche « → » permet de faire défiler très rapidement l'ensemble des documents d'un dossier.


**Figure 2.7.** Capture d'écran du mode « Coup d'œil » d'Apple.

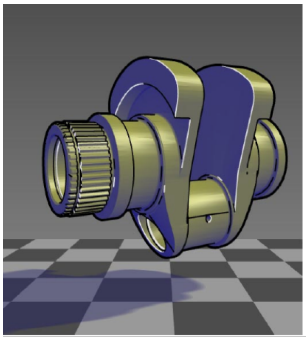
### 2.1.5 Méthodes de rendus non photoréalistes

Les techniques de ***Rendus Non Photoréalistes*** (NPR pour *Non Photorealistic Rendering*[2]), appelés aussi rendus expressifs, ne sont pas initiallement dédiées à la présentation de plusieurs documents et à l'exploration de bases de données. Toutefois nous avons jugé intéressant de les mentionner car, comme les stratégies de présentation vues précédemment, elles permettent de simplifier une image, ou d'accentuer certains éléments au détriment d'autres, afin de focaliser l'attention de l'utilisateur sur l'information à transmettre. Les techniques NPR sont couramment utilisées pour les dessins techniques, l'architecture, les illustrations et les schémas. Elles incluent également les techniques de stylisation qui permettent d'obtenir des rendus artistiques à partir de photographie (simulation d'aquarelles, de traits aux crayons...).

Par exemple, pour augmenter la compréhension du volume 3D d'un objet dans une illustration 3D, Gooch et ses collègues [Gooch *et al.* 1999] proposent d'accentuer l'information pertinente, ici l'indication de profondeur. Les auteurs manipulent les aspects visuels, comme l'épaisseur et la couleur des contours ou la teinte d'un objet, pour renforcer la perception de volume. Des exemples de rendus sont présentés figure 2.8.



**(a)** Accentuation des contours.



**(b)** Jeu sur la teinte chaude/froide.

**Figure 2.8.** Deux méthodes d'accentuation de la perception du volume 3D d'un objet dans une illustration. Extrait de [Gooch *et al.* 1999].

Au contraire, dans un schéma (dessin pour l'enseignement, notices...) on emploie plutôt la simplification voire la suppression de l'information non pertinente. Seule une partie du système

2. Ce nom a été choisi en opposition aux techniques photoréalistes qui cherchent à imiter le réel, c'est-à-dire à reproduire le comportement physique de la lumière dans une image de synthèse.

est représentée, et encore de manière simplifiée. Ces notions de simplification et de suppression sont très proches du concept d'affichage par filtrage de Furnas : on élimine ce qui est de moindre intérêt pour l'observateur. Par exemple, le schéma de l'oreille figure 2.9 [3] représente grossièrement chaque organe de l'oreille, les éléments plus fins que le niveau de détail souhaité n'apparaissent pas (par exemple on ne voit pas de représentations des cellules ciliées à l'intérieur de la cochlée), pas plus que le contexte (ici le reste de la tête humaine et en particulier le cerveau) qui reste masqué.

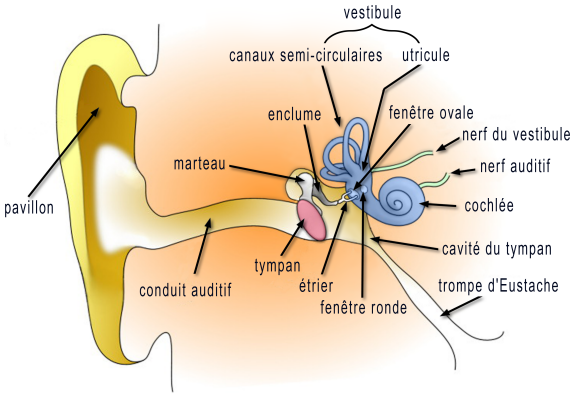



**Figure 2.9.** Exemplification de la suppression et de la simplification sur un schéma de l'oreille humaine. Seuls les éléments importants sont représentés.

## 2.2 Stratégies de présentation dans les interfaces sonores

Les stratégies de présentation que nous venons de décrire permettent de mettre en avant l'information pertinente en jouant sur le niveau de détail des différents documents ou parties de document. Cela rend possible l'affichage de plusieurs documents visuels simultanément et améliore la compréhension d'un document complexe qui comporte trop d'informations. Cependant ces stratégies reposent uniquement sur la modalité visuelle. Dans le contexte de la recherche de documents multimédia ayant une composante visuelle et une composante sonore, l'utilisateur peut avoir besoin des informations contenues dans la composante auditive. Nous allons voir qu'il existe des interfaces sonores (*auditory displays*) dédiées à l'accès à des informations auditives par leur présentation simultanée.

### 2.2.1 Intérêt spécifique de l'audio pour une interface homme-machines de recherche d'informations

Déjà été cités par les premiers chercheurs sur le sujet [Brewster 1994, p. 4-6], plusieurs points sont à prendre en compte pour la conception d'interfaces sonores :

- La modalité auditive peut apporter, dans certains cas, plus d'informations que la modalité visuelle. Par exemple, alors que la vision est limitée à un espace restreint en face de l'uti-

3. Issu de http://upload.wikimedia.org/wikipedia/commons/7/75/OreilleHumaine.png

lisateur, l'audition permet de percevoir des informations provenant de toutes les directions autour de lui.

- Certaines données peuvent également être représentées plus facilement par du son que par des images. Par exemple, l'évolution temporelle de l'intensité d'une onde sismique d'un tremblement de terre sera avantagesement sonifiée pour représenter l'importance du tremblement de terre [4].
- L'audition tient un rôle de système d'alarme et guide le regard dans une direction donnée en fonction de la position de la source sonore.
- Enfin pour des données sonores à l'origine (musique, émissions de radio, ...), il semble pertinent d'utiliser prioritairement cette modalité.

Pourtant les interfaces actuelles d'accès à des documents audio reposent majoritairement sur une représentation visuelle et/ou textuelle des documents. Ainsi un fichier de musique sera reconnu par une icone visuelle particulière liée à l'application qui permet d'ouvrir ce fichier, ou par l'extension (".wav", ".mp3") du nom de fichier. Bien que pratique dans une première approche, cette représentation ne donne aucune information sur le réel contenu sonore de ces documents et l'utilisateur devra écouter chaque document audio un par un.
Cette remarque vaut aussi pour les documents audiovisuels pour lesquels le contenu sonore est sous-exploité alors même qu'il est parfois plus pertinent que le document visuel (exemple d'un clip musical ou d'un documentaire avec commentaires en voix off). On peut aussi citer l'exemple des émissions de journaux télévisés : comme le cadre visuel varie peu lors de la présentation des sujets par le journaliste, seuls le contenu parlé (ici le son émis et le mouvement des lèvres) permet de distinguer deux sujets différents. A moins qu'une retranscription écrite par sous-titrage soit ajoutée (transformation de l'information sonore en information visuelle), l'information principale est portée par le son.

Une autre raison pour laquelle l'utilisation du son peut être recommandée est que la modalité visuelle est actuellement surexploitée. L'utilisateur peut être amené à exécuter plusieurs tâches en parallèle sur sa machine et ouvrir plusieurs fenêtres auxquelles il ne pourra pourtant pas accéder en même temps, compte tenu du manque d'espace écran. De plus, la surcharge cognitive qu'entraîne la présentation de trop nombreux éléments visuels peut induire l'utilisateur à manquer une information. En combinant audio et visuel, l'utilisateur pourrait alors percevoir et traiter plus d'informations qu'avec la modalité visuelle seule.

### 2.2.2 Interfaces sonores à base de présentation simultanée

En 1990 déjà, Ludwig et ses collègues [Ludwig *et al.* 1990] introduisaient le besoin d'établir des interfaces sonores avec des outils similaires à ceux disponibles dans le domaine visuel :

> *Users will soon need similar presentation, management, and organizational capabilities to avoid a confusing cacophony of multiple audio sources sounding at once.*

Selon eux les interfaces visuelles reposent sur la possibilité d'organiser spatialement des données affichées simultanément, et sur la capacité à changer son point d'attention d'un élément visuel

4. Un exemple de ce type de sonification est disponible à l'adresse http://www.youtube.com/watch?v=3PJxUPvz9Oo

à un autre. Ils revendiquent alors qu'il est possible d'organiser auditivement plusieurs sources sonores simultanées afin d'offrir des outils analogues aux outils visuels. Ils présentent un système de « fenêtrage » audio (en référence aux interfaces WIMP visuelles pour *Windows, Icons, Menus, Pointing device*) dans lesquelles les fenêtres en avant-plan sont plus proéminentes (plus fortes en volume sonore) que les fenêtres en arrière-plan. Plusieurs sources sonores sont donc jouées en même temps. Cependant, pour éviter la « cacophonie » induite par la présentation de ces nombreuses sources sonores simultanées, ils proposent de spatialiser les sons et d'introduire une hiérarchie entre les différents sons présentés. Il s'agit en fait de tirer profit des facultés du système auditif à séparer une scène sonore complexe en plusieurs flux auditifs [Bregman 1990] mais aussi à se concentrer sur une seule source sonore parmi les différentes sources présentées. Cette capacité, appelée ***effet cocktail party*** [Cherry 1953], est notamment augmentée par la perception de la position de chacune des sources qui sont séparées spatialement [Begault 1994]. Ainsi la spatialisation du son est-elle souvent exploitée dans les interfaces sonores d'exploration de bases de données sonores.

#### 2.2.2.1 Interfaces sonores zoomables

Parmi les ZUI sonores, l'***AudioStreamer*** [Schmandt et Mullins 1995] n'utilise pas de représentation visuelle et s'appuie donc uniquement sur le son. Elle permet d'écouter simultanément trois sources spatialisées virtuellement autour de l'utilisateur en trois positions distinctes. L'écoute se fait avec un casque sur lequel est monté un capteur d'orientation. L'utilisateur peut se tourner vers l'une des trois sources. Le système considère que l'utilisateur se tourne vers la source qui présente le plus d'intérêt pour lui. Le volume de cette source est alors augmenté de 10 dB par rapport au volume des autres sources de sorte qu'elle devienne acoustiquement prédominante. Il est alors plus facile pour l'utilisateur de séparer cette source des autres (ségrégation) et de l'écouter sélectivement. On retrouve ici les notions de focus (la source sonore prédominante) et de contexte (les autres sources).

Le projet ***Dynamic SoundScape*** [Kobayashi et Schmandt 1997] reprend ces concepts et les étend à un plus grand nombre de sources pour explorer auditivement le contenu d'un fichier sonore. L'idée de cette interface est de séparer un fichier sonore en plusieurs extraits et de placer virtuellement chacun de ces extraits à une position différente. La position temporelle de l'extrait dans le fichier originel est reliée à sa position spatiale autour de l'utilisateur. Ainsi ce dernier peut écouter simultanément plusieurs passages d'un même fichier sonore. Comme dans l'***AudioStreamer***, l'auditeur équipé d'un système de *tracking* (capteur d'orientation) peut se tourner sur lui-même. L'écoute sélective est renforcée par une adaptation du niveau sonore selon l'orientation de l'auditeur. Le volume dépend alors de la distance angulaire entre la direction de la tête de l'utilisateur et la position allouée à l'extrait sonore. L'extrait face à l'auditeur est plus fort de 8 dB par rapport aux extraits situés derrière l'utilisateur. Là encore on retrouve les notions de focus+contexte, avec en plus un parallèle entre la déformation progressive du volume sonore et la déformation visuelle de la taille des objets dans une technique de lentille grossissante (*Fisheye*).

Une autre application de Chris Schmandt pour l'exploration de bases de données audio est le ***Audio Hallway*** [Schmandt 1998]. Cette application exploite la métaphore du couloir et de

pièces dans lesquelles sont rangés les sons et que l'on peut ouvrir. Cela permet ainsi d'obtenir plusieurs niveaux de détails pour un ensemble de sons. En effet, des extraits radiophoniques sont regroupés en fonction de leur méta-données puis mixés grossièrement avant d'être placés dans des « salles virtuelles ». L'utilisateur peut alors parcourir le couloir virtuel et percevoir les sons mixés qui émanent de chaque « salle ». Le rendu perçu est donc un mélange peu détaillé qui permet d'avoir une vue d'ensemble de la collection audio. Le niveau sonore de chaque salle varie en fonction de la distance de l'utilisateur par rapport à cette salle dans l'environnement virtuel. En « entrant » dans une salle, l'utilisateur entend uniquement les sons du groupe de cette salle, cette fois séparément au lieu d'être mixés. Les mixages des autres salles sont rendus silencieux. Les sons de la salle « visitée » sont donc plus distincts.

Dans les applications ***SonicBrowser*** [Fernström et Brazil 2001; Brazil 2003] ) et ***SoundTorch*** [Heise *et al.* 2008, vidéo 2.1], l'exploration de la collection se fait en naviguant dans un espace sonore 2D. Pour la diffusion des sons, les deux applications utilisent un même concept d'*aura* (ou *torche* dans ***SoundTorch***). Il s'agit d'un disque dans une représentation visuelle 2D de la collection, qui définit virtuellement la limite de perception auditive de l'utilisateur : tous les objets à l'extérieur du disque sont silencieux (niveau sonore nul), tandis que les objets situés à l'intérieur du disque sont diffusés avec un niveau sonore relatif à leur distance par rapport au centre du cercle, les plus proches étant plus forts en volume. Les sons sont spatialisés autour de l'auditeur de telle manière que leur position perçue corresponde à leur place par rapport à l'auditeur lorsque l'on regarde l'espace 2D du dessus (voir figures 2.10 et 2.11). Au contraire des interfaces précédentes, ces deux applications reposent en plus sur une représentation visuelle des données sonores sous forme d'icônes visuelles. Elles exploitent ainsi de nombreuses stratégies de présentation visuelle des données telles que les graphes 2D x-y et renseignent sur la structure de la base de données (arbres hiérarchiques) [Brazil 2003]. Cependant il ne s'agit que d'une visualisation de données sonores et non de l'affichage d'informations visuelles contenues au préalable dans des données multimédia. Ainsi ces applications ne sont pas adaptées à l'exploration de bases de données multimédia.

2.1

Ces deux dernières interfaces ont été explicitement inspirées des techniques de présentation visuelle en particulier du concept de lentille grossissante (ou *Fisheye*). On retrouve alors la notion de facteur d'échelle qui définit ici une sorte de ***zoom audio***. En effet, pour les deux derniers exemples ***SonicBrowser*** et ***SoundTorch*** en jouant sur le rayon de l'aura, on peut ajuster un facteur de zoom de sorte que si le rayon est très grand on peut entendre une grande quantité de sons (zoom arrière) tandis que si le rayon est faible on peut avoir un « gros plan » sur un petit nombre de sons (zoom avant).

#### 2.2.2.2 Limites des interfaces sonores à base de simultanéité et enjeux

Les premières limites des interfaces sonores simultanées telles que nous venons de les présenter dans la section précédente dépendent de la modalité audio en elle-même. En effet la mise en œuvre des interfaces sonores est rendue plus difficile que celles des interfaces visuelles du fait que :

- la séparation de sources sonores simultanées est plus complexe que pour le visuel ;

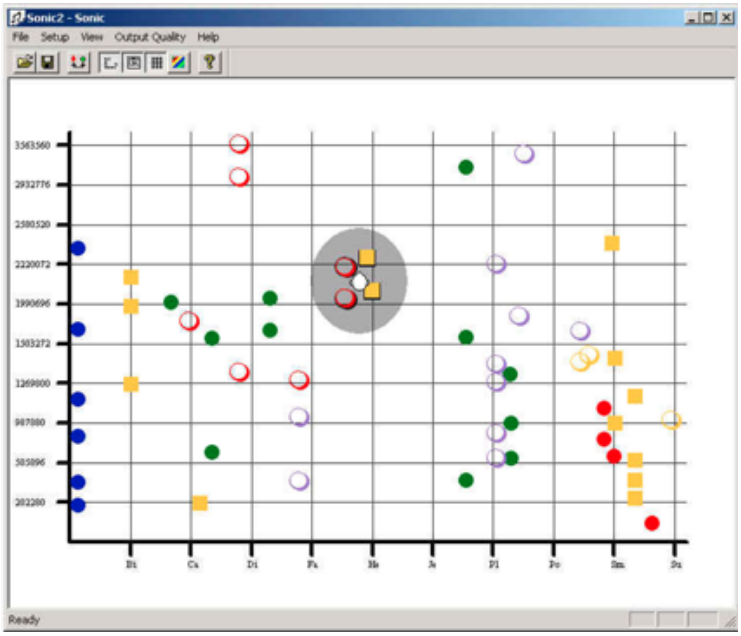



**Figure 2.10.** Capture d'écran du ***SonicBrowser*** version 2. La carte 2D représente le plan horizontal d'écoute (vue du dessus). Le disque gris est l'*aura*. Seuls les sons dont l'icône est située dans ce cercle seront joués. Le centre du disque est associé à la tête de l'utilisateur, les sons sont spatialisés en fonction de leur position par rapport à l'utilisateur dans le plan 2D. Extrait de [Fernström et Brazil 2001].

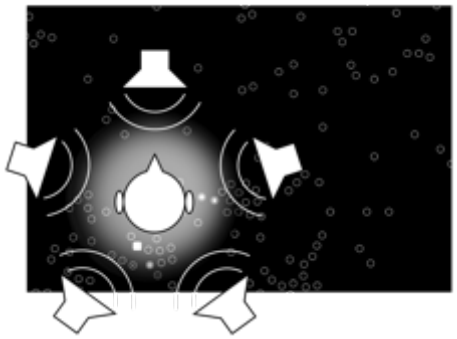


**Figure 2.11.** Capture d'écran du rendu du ***SoundTorch*** et représentation schématique des différents haut-parleurs virtuels. Extrait de [Heise *et al.* 2008] .

- nos connaissances sur la perception auditive et la perception audiovisuelle sont encore restreintes si on les compare à nos connaissances sur le système visuel humain ;
- l'audio est spécifiquement dynamique et change donc en continu alors que les interfaces visuelles actuelles (à l'exception de quelques méthodes dédiées à la vidéo) reposent sur une présentation statique à partir d'images fixes.

Cette modalité présente d'autres inconvénients par rapport au visuel, mais qui ne nous concerneront pas directement dans notre projet d'exploration individuelle de collections multi-

média. En effet, l'utilisation de l'audio dans une interface destinée à une utilisation collective, où plusieurs utilisateurs utilisent le système en même temps et peuvent être amenés à demander une présentation d'informations différentes, peut être contraignante du fait que la diffusion sonore est moins facile à limiter dans l'espace que la diffusion visuelle (un écran chacun par exemple). Une solution est alors de fournir à chaque utilisateur un casque pour qu'il puisse bénéficier du rendu audio qui lui est destiné sans perturber les autres utilisateurs. Cette méthode est alors plus intrusive et a l'inconvénient de couper les utilisateurs du monde extérieur, les empêchant alors de communiquer directement.

De plus, l'apparition de moyens de reproduction et de stockage de sons sur des supports a été tardive par rapport aux techniques de reproduction visuelle (peinture, photographie). En informatique aussi, l'utilisation de l'audio est beaucoup plus récente que celle du visuel. Par conséquent, le domaine des interfaces sonores est plus jeune que celui des interfaces visuelles. Enfin, plusieurs questions restent en suspens concernant les interfaces sonores basées sur la présentation simultanée de plusieurs sons [McGookin et Brewster 2006]. Si l'idée de ces interfaces date d'il y a environ 15 ans [Schmandt et Mullins 1995], trop peu de tentatives ont été menées depuis.

### 2.2.3 Autres méthodes sonores applicables : lecture rapide

Les interfaces sonores ne reposent pas toutes sur la simultanéité et l'on trouve d'autres méthodes qui permettent d'accéder à un maximum d'informations, ou à l'information pertinente, en un minimum de temps.

Certaines d'entre elles rendent l'accès à l'information sonore plus rapide par une lecture accélérée des sons. Cette approche est donc très proche de la technique RSVP visuelle (*Rapid Serial Visual Presentation*). Par exemple, les ***spearcons*** permettent d'accélérer la lecture des icônes sonores : un mot est enregistré puis accéléré [Walker *et al.* 2006]. Des exemples de *spearcons* sont présentés dans les exemples audio 2.1 et audio 2.2 [5]. Une autre technique, appelée ***audio skimming***, permet de parcourir très rapidement un fichier sonore en accélérant ou ralentissant la lecture du signal audio tout en limitant la distorsion apparente [Couvreur *et al.* 2008]. L'utilisateur peut alors trouver plus rapidement le segment du fichier qui l'intéresse. L'exemple audio audio 2.3 [6] illustre cette technique sur un extrait musical : la vitesse de lecture est accélérée ou ralentie selon les besoins de l'utilisateur.

2.1 2.2 2.3

## 2.3 Stratégies de présentation dans les interfaces multimédia

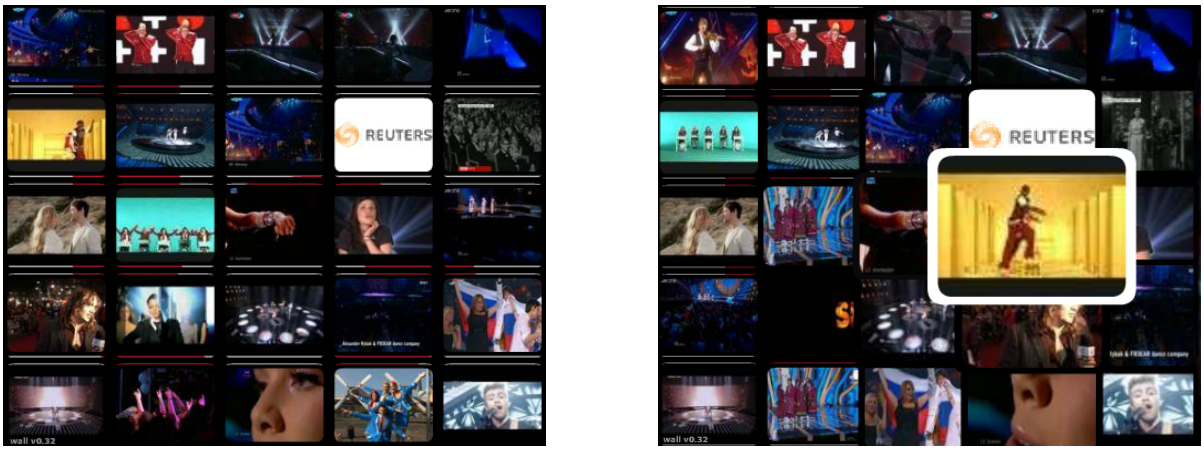
Nous avons exposé des méthodes de présentation pour le visuel seul ainsi que des méthodes pour l'audio seul. Nous allons maintenant discuter les recherches sur les interfaces zoomables et les méthodes dédiées à la présentation de documents audiovisuels, en particulier de vidéos.

5. Exemples de *spearcons* issus de http://sonify.psych.gatech.edu/research/auditorymenus/index.html
6. Son extrait de la vidéo 1 du projet http://www.numediart.org/projects/01-1-audio-skimming/

### 2.3.1 Stratégies visuelles de présentation de vidéos

La plupart des outils actuels de présentation vidéos dans de grandes collections, disponibles, par exemple, pour les applications de vidéos à la demande (*video-on-demand*) comme *YouTube*, *DailyMotion*, *Google Videos*, et d'autres reposent sur une présentation visuelle de type Overview+Detail : de nombreuses vignettes, sous forme d'images fixes, sont disposées sur un bord de l'écran accompagnées de données textuelles tandis qu'une seule vidéo peut être lue avec le son. Cette méthode d'affichage oblige l'utilisateur à se baser essentiellement sur l'aspect graphique. En revanche, s'il n'a comme critère de recherche qu'un critère auditif, il est contraint de cliquer sur chacune des vidéos une par une jusqu'à trouver ce qu'il cherche. De plus, l'affichage de vignettes implique une pré-sélection d'une image-clé, censée représenter au mieux l'ensemble du contenu vidéo (quand il ne s'agit pas simplement de la première image de la vidéo), et ce pour chaque vidéo. Il se peut que cette image-clé ne corresponde pas à l'idée que se fait l'utilisateur de la vidéo, par exemple si l'utilisateur en a déjà vu le contenu mais se souvient d'une scène ou d'un personnage ne correspondant pas à l'image-clé. Dans ce cas, l'utilisateur aura plus de difficultés à trouver la vidéo qu'il recherche et sera, là encore, amené à cliquer sur chacune des vidéos une par une.

L'affichage proposé dans le mode « Wall » du moteur ***BlinkX*** résout partiellement ce problème puisqu'on peut accéder au contenu vidéo de plusieurs vidéos simultanément. Sur la page d'accueil plusieurs vidéos, et non leur image-clé respective, sont affichées en grille et sont jouées simultanément (figure 2.12a). De plus, le curseur de la souris permet d'afficher une des vidéos plus grosse que les autres. Cette vidéo passe alors légèrement au dessus des autres et la taille des autres vidéos diminue progressivement avec l'éloignement par rapport à la vidéo grossie (figure 2.12b). Il s'agit donc d'une technique mixte entre la juxtaposition masquante de l'affichage bifocal et la vue déformée de la lentille grossissante en œil-de-poisson. Le son des vidéos ne peut pas être entendu à moins de sélectionner une des vidéos du mur par un clic. La lecture de l'audio se fait donc séquentiellement, une vidéo à la fois.




**(a)** Curseur à l'extérieur de la fenêtre. **(b)** Déformation sous le curseur.

**Figure 2.12.** Captures d'écran du *BlinkX Wall* pour la requête « Eurovision ». La vidéo sous le curseur est grossie, les vidéos autour sont réagencées.

Une autre technique appelée ***Fisheye Videos*** avait été proposé dès 1993 par Yamaashi et ses collègues pour la présentation de plusieurs vidéos simultanées [Yamaashi *et al.* 1993]. Elle répondait surtout au problème de l'époque concernant les limites des ressources de calcul et de mise en mémoire de nombreuses vidéos. La solution proposée consiste à adapter, à la fois la résolution spatiale des images et la cadence de lecture (le nombre d'images par seconde) de chaque vidéo. Les différentes vidéos sont affichées dans des fenêtres distinctes que l'utilisateur peut superposer. Les deux paramètres de cadence et résolution spatiale sont alors calculés en fonction de la position de la fenêtre (une fenêtre au dessus des autres est considérée comme plus intéressante) et du pourcentage de zone masquée de la vidéo (une fenêtre entièrement visible est considérée comme plus intéressante) : plus le degré d'intérêt de la vidéo est petit et moins la résolution spatiale et la cadence seront élevées. Cette méthode présente une lacune importante car elle ne tient pas compte de la composante sonore des vidéos. De plus, comme le prouve le « Wall » de BlinkX, les ressources disponibles actuellement sur nos machines nous permettent, dans une certaine mesure, d'afficher plusieurs vidéos simultanément sans avoir besoin de diminuer ni la cadence ni la résolution spatiale des documents.

### 2.3.2 Avantages de la multimodalité

Les deux méthodes d'affichage de vidéos présentées juste avant ne proposent qu'un affichage de la composante visuelle des vidéos et ne tiennent pas compte de la composante audio. Il s'agit donc de méthodes unimodales. Pourtant, l'utilisation de la multimodalité a souvent montré qu'elle pouvait entrainer de meilleures performances et ce dans de nombreuses tâches. Parmi les modalités sur lesquelles nous nous sommes penchée, les modalités auditives et visuelles ont montré qu'elles pouvaient être combinées de façon très bénéfique. D'un point de vue purement perceptif, leur combinaison peut diminuer le temps nécessaire pour l'identification d'objet si l'information portée par les deux modalités n'est pas incohérente (voir chapitre 3, section 3.4, page 47). On parle d'effet de ***facilitation auditive*** lorsque l'audio devient un soutien pour une interface visuelle. De la même manière le visuel peut augmenter une interface sonore. On parle de ***facilitation visuelle*** [Boyne *et al.* 2005].

De plus, l'audio peut être ajouté de façon redondante par rapport à l'affichage graphique. Procédant ainsi, l'ajout d'audio n'est pas intrusif. Il ne modifie pas les méthodes d'interaction visuelles et n'interfère pas avec les habitudes de l'utilisateur. La plupart du temps, les utilisateurs n'éprouvent d'ailleurs pas de gêne lorsque confrontés à des interfaces sonores et en sont au contraire soulagés. Toutefois si l'utilisateur n'est pas convaincu par l'ajout de cette modalité, il lui suffit de la rendre silencieuse alors qu'il ne peut se passer de l'information visuelle.

Finalement, l'utilisation d'une autre modalité est nécessaire pour s'adapter à certains utilisateurs, par exemple les non ou mal-voyants, ou dans le cas où la modalité visuelle est déjà employée pour d'autres tâches (typiquement la conduite) auquel cas l'utilisateur ne peut pas porter toute son attention sur un support visuel. La modalité auditive a alors montré être un bon substitut pour ce type d'utilisation.

### 2.3.3 Stratégies de présentation audiovisuelles

Nous n'avons connaissance que d'une seule méthode de Focus+Contexte multimodale. Il s'agit de l'application ***Dolphin*** dans laquelle une combinaison audiovisuelle permet d'afficher simultanément plusieurs informations cartographiques (restaurants, attractions...) pour augmenter une application GPS [McGookin et Brewster 2002]. Les auteurs se basent sur la capacité du système auditif humain à percevoir les informations à 360°, ce qui permet d'obtenir des informations globales. Au contraire le système visuel humain sert surtout à capter des détails mais dans un champ visuel très restreint pour sa zone de focus. Ainsi la présentation de ***Dolphin*** utilise l'audio pour donner des informations sur le contexte tandis qu'une représentation détaillée est disponible par affichage visuel sur un écran d'appareil mobile. On peut voir cela comme une technique d'Overview+Detail dans laquelle la vue d'ensemble et le détail sont séparés, non plus dans deux fenêtres visuelles distinctes, mais dans deux modalités distinctes. Toutefois les résultats d'une expérimentation menée par les auteurs ne sont pas très encourageants. En effet, aucune différence n'est observable entre cette technique multimodale et une présentation visuelle seule. Les auteurs suggèrent que les gains apportés par l'audio sont contrebalancés par la présence de trop nombreuses sources sonores de même niveau, qui fatiguent l'auditeur ou induisent de mauvaises interprétations. On peut se demander si le fait que les différentes sources sonores ne soient pas hiérarchisées mais toutes affichées sur un même plan n'est pas à l'origine des inconvénients de cette méthode. Dans les études menées durant cette thèse, nous avons justement étudié comment une stratégie de distorsion auditive peut être combinée à une présentation visuelle et si cela diminue l'éventuelle gène sonore. Quoi qu'il en soit, compte tenu que cette technique s'appuie sur une séparation de l'information en deux modalités, et ce sans redondance, elle n'est pas adaptée à la présentation de vidéos.

## 2.4 Conclusion

L'objet de ce chapitre était de faire un état des lieux des méthodes de présentation exploitables à l'heure actuelle pour présenter des documents visuels, audio ou multimédia.

Nous avons pu remarquer que de nombreuses méthodes ont été conçues pour des interfaces visuelles, tandis que seules quelques unes ont été développées pour des interfaces sonores. De plus la grande majorité des stratégies auditives a d'abord été inspirée par des stratégies visuelles (exemple : la torche sonore inspirée de la lentille en œil-de-poisson), et pour les autres, il est possible de trouver un analogue dans la modalité visuelle (exemple : la présentation sérielle rapide a été développée indépendamment dans les deux modalités). En outre, le nombre de paramètres visuels employés dans les ZUI visuelles est plus importants que le nombre de paramètres sonores employés dans les ZUI audio. On peut alors se demander s'il est possible d'augmenter le nombre de paramètres sonores utiles pour la présentation d'informations auditives. Nous étudierons comment étendre à la modalité auditive les paramètres visuels non encore exploités en audio, en particulier les notions de profondeur de champ sémantique et de transparence.

De plus, puisque les méthodes de présentation destinées au browsing de vidéos reposent de façon quasi exclusive sur la modalité visuelle, nous nous proposons de tester des combinaisons de méthodes visuelles présentées auparavant et de méthodes auditives soit présentées auparavant (chapitre 4) soit qu'on aura obtenues comme analogues de méthodes visuelles (chapitre 6).

Enfin, dans le grand public, peu d'interfaces de browsing audio sont basées sur une lecture simultanée des sources sonores. Les raisons de ce manque d'interfaces sonores ne sont pas bien définies et l'on peut se demander si cela est dû à un problème technique ou plutôt aux limites perceptives lors de l'écoute de plusieurs sons concurrents. Nous chercherons donc à répondre aux questions suivantes : dans quelle mesure peut-on exploiter la simultanéité en audio ? Quand atteint-on une surcharge cognitive, c'est-à-dire quand sature-t-on les sens de l'utilisateur ? Cela nous fait nous interroger sur les principes de perception sous-jacents aux tâches de recherche et d'exploration de plusieurs documents simultanés. On s'intéressera donc à la perception visuelle, sonore, mais aussi multisensorielle que nous présentons dans le chapitre suivant.

Chapitre 3

# Analyse des recherches en perception auditive, visuelle et multisensorielle appliquées à la recherche de document

*Ce que je vois et ce que j'entends du monde extérieur,*
*c'est simplement ce que mes sens en extraient pour éclairer ma conduite.*

Henri Bergson, *Le Rire*

## Sommaire



Les interfaces les plus faciles à utiliser pour l'utilisateur sont celles qui sont bien adaptées à ses besoins, à ses connaissances et à ses facultés, comme le soulignent Rosenbaum et ses collègues [Rosenbaum et Schumann 2009] :

> *Intuitive user interfaces significantly increase usability and acceptance of browsing techniques. An intuitive interface is aligned to human capabilities and does not require extensive learning.*

Nous nous sommes alors intéressée aux processus attentionnels et perceptifs que les utilisateurs mettent en place lors de l'exploration de bases de données multimédia et de la recherche d'un document spécifique. Le chapitre qui suit présente les facteurs humains mis en jeu dans l'identification, la détection et la recherche d'un stimulus audiovisuel présenté parmi plusieurs stimuli concurrents. En particulier, nous nous sommes concentrée sur la notion d'attention qui peut être définie, d'après William James [James 1890], comme *la prise de possessions par l'esprit sous une forme claire et vive d'un objet ou d'une suite de pensées parmi plusieurs possibilités [...]. Elle implique le retrait de certains objets afin de traiter plus efficacement les autres*[1]. Le lecteur pourra également se référer à l'ouvrage de Pashler [Pashler 1998] qui offre un état de l'art sur les phénomènes perceptifs et attentionnels sous toutes leurs formes et évoque donc les aspects visuels, sonores et multisensoriels que nous abordons ci-après.

Ce chapitre est organisé en trois sections principales : la première partie expose les attributs perceptifs mis en jeu dans les tâches de recherche visuelle, la seconde partie ceux mis en jeux dans les tâches de recherche sonore, tandis que la troisième partie s'intéresse aux interactions entre les deux sens, c'est-à-dire à l'intégration multisensorielle.

## 3.1 Tâche de recherche : perception et attention en visuel

En général, lorsqu'un utilisateur recherche un document dans une grande base de données, plusieurs documents lui sont présentés. Nous avons vu au chapitre 2 que la plupart des techniques de présentation utilise une présentation simultanée. Lorsque l'utilisateur sait exactement quel document il recherche, on emploie le terme ***searching***[2]. L'utilisateur doit retrouver le document qui l'intéresse, la ***cible***, parmi plusieurs documents de moindre intérêt, les ***distracteurs*** qui lui sont présentés en même temps.

En visuel, cette tâche de recherche (***visual search***) a été fortement documentée, notamment depuis l'introduction par Neisser en 1967 de la notion de traitement préattentif de l'information [Treisman et Gelade 1980; Treisman 1986; Healey et Enns 2011; Ware 2000, chapitre 5]. On parle de ***préattention*** pour désigner les étapes perceptives précoces supposées avoir lieu avant que n'intervienne l'attention volontaire et consciente. Pour un ordre d'idée, toute tâche qui peut être réalisée, sur une présentation de nombreux éléments simultanés, en moins de 200 ou 250 msec, est considérée par la communauté scientifique comme préattentive.

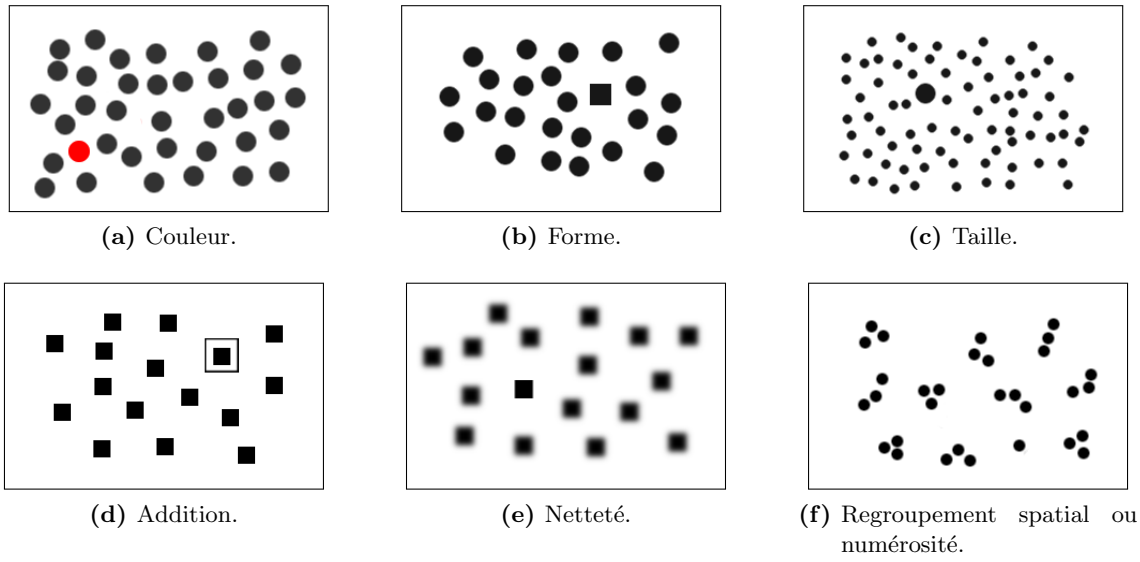
Pour illustrer cette notion, nous pouvons prendre l'exemple d'une personne au milieu d'une foule. Si cette personne porte un pull rouge tandis que les autres portent un costume noir, alors elle attire l'œil et son pull « saute aux yeux ». On parle de phénomène de ***pop-out***. La capacité d'un objet à émerger des autres est appelée ***saillance***. Les paramètres visuels qui, comme la couleur, permettent de faire ressortir un objet extrêmement rapidement sont communément qualifiés de préattentifs (*preattentive features*).[3]

1. Traduit de *[Attention is] the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thoughts. [...] It implies withdrawal from some things in order to deal effectively with others.*

2. Au contraire, quant l'utilisateur ne sait pas précisément ce qu'il recherche, il est obligé de survoler la collection. On parle alors de *browsing*.

3. Ces paramètres ont été initiallement qualifiés de ***préattentifs*** mais nous savons aujourd'hui que l'attention joue un rôle important dans le traitement de l'information visuelle même dans cette étape précoce de la vision.

Nous préfèrerons parler de ***traits caractéristiques***. Un élément visuel qui possède un trait caractéristique est naturellement plus saillant que les autres et aura donc une très forte probabilité d'être remarqué. Un trait caractéristique est donc un paramètre visuel que la cible ne partage pas avec les distracteurs. Différentes dimensions (paramètres visuels) peuvent faire la distinction entre cible/distracteurs, certaines étant plus intéressantes en visualisation d'information car elles sont traitées en grande partie de façon involontaire et donc sans effort pour l'utilisateur. Des exemples de ces dimensions sont présentés en figure 3.1. Elles peuvent être regroupées en plusieurs catégories principales comme la forme, la couleur, la position et le mouvement [Ware 2000, chapitre 5, p. 163–170]. Une liste non exhaustive de traits caractéristiques est présentée par catégorie dans le tableau 3.1.



**(a)** Couleur. **(b)** Forme. **(c)** Taille.

**(d)** Addition. **(e)** Netteté. **(f)** Regroupement spatial ou numérosité.

**Figure 3.1.** Exemples de traits caractéristiques : les objets qui diffèrent attirent automatiquement l'attention.

### 3.1.1 Combinaison de traits caractéristiques

Le traitement involontaire et rapide de ces traits caractériques (phénomène de pop-out) ne concerne que les cas où la cible ne varie des distracteurs que par une seule dimension, ou plusieurs dimensions redondantes, que les distracteurs ne partagent pas (e.g. une cible cercle rouge au milieu de distracteurs carrés noirs). Cette tâche est alors appelée ***feature search***. Dans ce cas, le temps nécessaire pour retrouver la cible ne dépend pas du nombre de distracteurs. Pour reprendre notre exemple, retrouver une personne au pull rouge au milieu d'une foule prendra le même temps que la foule compte 10 personnes ou 100, du moment que toutes les

Le terme de *preattentive features* est cependant resté car il fait intuitivement référence à la grande rapidité avec laquelle ces paramètres sont détectés ($< 250$ *msec*). Nous employons ici le terme de ***traits caractéristiques*** qui permet de regrouper à la fois les paramètres préattentifs simples mais aussi leur combinaisons non redondantes qui, bien que détectés plus lentement, restent des paramètres aisément reconnaissables qui peuvent augmenter la saillance des objets.

| Forme | Couleur | Mouvement | Position spatiale |
|---|---|---|---|
| Taille | Teinte | Direction | Position 2D |
| Longueur de traits | Saturation | | Profondeur |
| Largeur de traits | Clignotement | Vitesse | Orientation 3D |
| Orientation de ligne(s) | | | Renfoncement convexe /concave |
| Intersection | | | |
| Courbure | | | Regroupement spatial |
| Densité | | | |
| Netteté | | | |

**Table 3.1.** Exemples de traits caractéristiques visuels par catégorie.

autres personnes de la foule sont habillées uniformément de la même couleur distincte du rouge. Dans le cas de dimensions redondantes, c'est-à-dire lorsque la cible se distingue des distracteurs par deux traits caractériques, le temps de recherche pour retrouver la cible donne des temps de recherche encore plus courts. Cet avantage est appelé ***redondance de cible*** (*redundant targets*) [Krummenacher *et al.* 2001; Grubert *et al.* 2011].

Cependant nous sommes rarement confrontés à ce type de recherche et d'autres facteurs perceptifs et attentionnels entrent alors en jeux. Ainsi lors d'une recherche dite de ***conjonction*** (*conjunction search*), la cible est définie par une combinaison d'au moins deux traits caractéristiques, comme un objet de forme cercle et de couleur bleue, alors qu'il y a plusieurs types de distracteurs qui possèdent chacun une seule de ces caractéristiques, des cercles rouges, des carrés rouges et des carrés bleus. Dans ce cas, la cible n'est pas aussi facilement détectée et l'effet de pop-out disparaît (voir figure 3.2). Les traitements préattentifs permettent cependant de limiter, dans un premier temps, l'espace de recherche avant de procéder à une étape de *scanning* où les objets préselectionnés sont analysés séquentiellement (et non plus en parallèle) jusqu'à trouver la cible. Ainsi pour reprendre l'exemple, on traitera uniquement les objets bleus, la distinction entre objets bleus/ objets rouges étant préattentive, et l'on pourra parcourir un par un les objets du sous-espace « objets bleus ». Cette tâche nécessite donc un effort de la part de l'utilisateur qui devra volontairement porter son attention sur chacun des objets successivement.

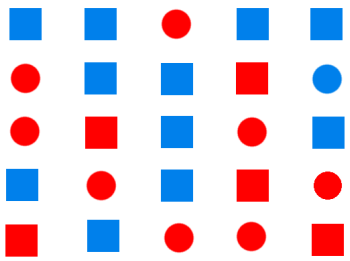


**Figure 3.2.** Exemple de recherche d'une cible à deux traits caractéristiques : ici un rond bleu.

### 3.1.2 Hiérarchie entre traits caractéristiques et asymétrie

Lorsque plusieurs traits caractéristiques entrent en jeu, il peut y avoir des effets d'interférence. En effet, certains paramètres visuels prédominent. On parle de hiérarchie entre les paramètres. Par exemple, il est difficile d'ignorer les propriétés de couleur et de distinguer les objets entre eux uniquement en fonction de leur forme. Il est intéressant de noter qu'il y a alors asymétrie entre les paramètres et le paramètre forme n'a aucune influence sur notre capacité à séparer les objets selon leur couleur [Ware 2000, *2nde édition*, chapitre 5, p. 153]. Un autre exemple d'asymétrie se retrouve dans le problème de la familiarité cible/distracteurs. Ainsi il est plus facile de retrouver une cible inconnue dans un ensemble de distracteurs connus qu'une cible connue dans un ensemble de distracteurs inconnus [Wang *et al.* 1994].

Lors d'un processus cherchant à mettre en valeur un élément (*highlighting*), nous retiendrons surtout qu'il faut en général mieux ajouter une marque à la cible et laisser les distracteurs inchangés que faire l'inverse. Ainsi il faut mieux souligner ce qui est important que tout ce qui ne l'est pas [Wolfe 2001; Treisman et Gormican 1988].

Enfin, compte tenu de la hiérarchie entre les paramètres et des interférences possibles entre ces paramètres, il faut éviter au maximum que les distracteurs aient des traits caractéristiques propres selon d'autres paramètres que celui utilisé pour la mise en valeur. Ainsi, si l'on utilise la taille pour mettre en valeur une vidéo cible plus grande que les autres, mais qu'une des autres vidéos est l'unique vidéo dans les tons rouges, on aura sans doute plus de difficultés à attirer l'attention de l'utilisateur sur la vidéo cible.

### 3.1.3 Facteurs limitants

D'autres facteurs, indépendants des propriétés visuelles de la cible elle-même, rendent la distinction entre la cible et les distracteurs plus ou moins difficile [Duncan et Humphreys 1989] :

- Le **nombre de distracteurs** (*set-size effect*) agit sur la recherche en conjonction : plus le nombre de distracteurs augmente et plus la tâche est difficile ;
- La **similarité entre la cible et les distracteurs** : plus la similarité est forte et plus la tâche est difficile (retrouver un objet rouge au milieu d'objets roses est plus dur que de retrouver cet objet au milieu d'objets bleus) ;
- L'**homogénéité des distracteurs** (ou similarité distracteur-distracteurs) sur la propriété visuelle qui sert de trait caractéristique : plus les distracteurs sont hétérogènes et plus il sera difficile de retrouver la cible. Cela arrive si pour une même propriété visuelle on a différents niveaux de valeur possible (par exemple rechercher un objet rouge parmi des objets de toutes les couleurs) ;
- La **familiarité et mémorisation de la cible/des distracteurs** : nous avons déjà mentionné l'asymétrie présentée par [Wang *et al.* 1994] concernant la plus grande facilité à retrouver une cible inconnue dans un ensemble de distracteurs connus qu'une cible connue dans un ensemble de distracteurs inconnus. En fait, plus récemment Shen et Reingold [Shen et Reingold 2001] ont montré que la familiarité de la cible n'a pas d'influence, seule celle des distracteurs en a : le temps de réponse raccourcit quand les distracteurs présentés sont connus. Ainsi la recherche d'un document visuel dans une collection de documents connus (une collection personnelle par exemple) ira plus vite que si la collection n'est pas connue ;

- Le **nombre de cibles** à trouver : plus le nombre de cibles est élevé, plus le temps de recherche est long ;
- L'**excentricité de la cible** : le champ visuel se décompose en deux parties, d'une part le champ visuel central où la précision visuelle est élevée et le champ visuel périphérique de faible résolution. Repérerer une cible, même saillante, située dans le champ visuel périphérique est plus difficile que dans le champ visuel central ;
- La **présence d'un stimulus d'une autre modalité** : par exemple, la présence d'un son indiquant la direction de la cible visuelle [Simpson *et al.* 2010], ou synchronisé avec la cible [Van der Burg *et al.* 2008], ou encore en lien sémantique avec elle [Iordanescu *et al.* 2010 2011] accélère la recherche. Ce point sera abordé plus en profondeur dans une section spécifique sur la perception multisensorielle (section 3.4, page 51).

### 3.1.4 Utilisation de la préattention visuelle dans les IHM

L'utilisation de la saillance perceptive et de l'effet de *pop-out* visuel est fréquent dans les IHM, notamment pour la présentation ou la visualisation de données [Healey *et al.* 1995]. Ware consacre d'ailleurs tout un livre sur comment appliquer la perception à la visualisation et au design [Ware 2000].

Différentes tâches de visualisation peuvent ainsi bénéficier des avantages de la saillance perceptive comme [Healey *et al.* 1995; Kosara 2001] :

- la recherche visuelle/la recherche de documents ;
- la mise en relief/hiérarchisation d'éléments ;
- l'estimation du nombre d'éléments appartenant à un groupe ;
- le regroupement d'éléments ;
- l'apprentissage.

Intéressée plus particulièrement par la recherche de documents et la mise en relief/hiérarchisation d'éléments, nous retiendrons surtout l'utilisation très répandue de ces traits caractéristiques dans les méthodes focus+contexte, comme on a pu le voir au chapitre 2, section 2.1, :

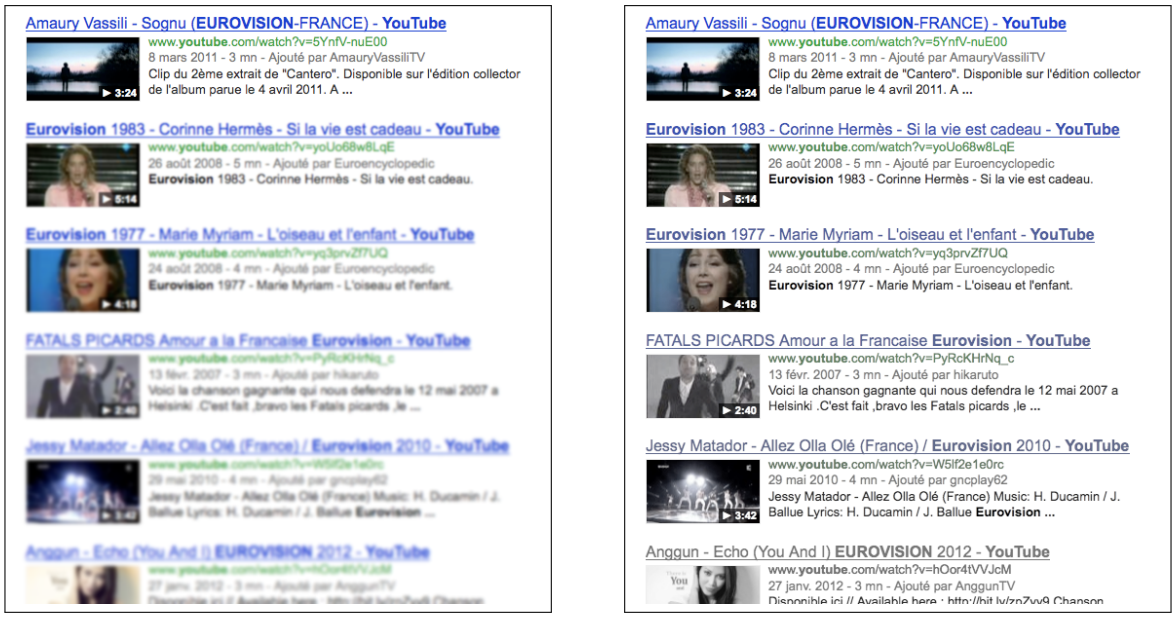
- le paramètre de netteté flou est utilisé dans la méthode de profondeur de champ sémantique (SDOF pour *Semantic Depth of Field*) [Kosara *et al.* 2002ba] ;
- la taille est utilisée dans les techniques d'affichage bifocal [Apperley et Spence 1982] et de lentille en œil-de-poisson [Sarkar et Brown 1994] ;
- la couleur permet de renforcer les distorsions sur la taille dans les lentilles grossissantes sous forme d'ombre [Zanella *et al.* 2000], ou les distorsions de flous par de la désaturation [Rosenbaum et Schumann 2009] ;
- Suh et ses collègues [Suh *et al.* 2002] proposent d'utiliser une combinaison taille/couleur pour améliorer les interfaces Overview+Detail.

Enfin, d'autres méthodes de visualisation, comme les techniques NPR (non photoréalistes) (page 18), utilisent certains paramètres préattentifs pour diriger le regard sur l'information importante, par exemple la teinte et le contraste renforcent les aspects de volume sur les schémas de dessin technique.

## 3.2 Proposition d'application de la préattention au contexte de la recherche de vidéos

Dans les interfaces d'exploration multimédia actuelles, plusieurs documents visuels réalistes, photographies ou vidéos, sont présentées simultanément côte à côte. Leur contenu est souvent très varié et ils sont donc très dissimilaires. La cible ne possède alors pas de trait caractéristique suffisamment saillant : l'utilisateur est amené à faire une recherche sérielle pour retrouver le document qui l'intéresse. Nous nous proposons d'exploiter les traits caractéristiques et facteurs préattentifs que nous venons d'exposer et de les intégrer dans des outils de présentation et de navigation dédiés aux collections multimédia.

Une première idée est de hiérarchiser l'affichage des différents documents selon leur intérêt supposé pour l'utilisateur. Cette utilisation serait destinée au cas où l'utilisateur a une connaissance préalable de ce qu'il recherche. Par exemple, les réponses proposées par un moteur de recherche sont classées par taux de similarité avec la requête. En jouant sur les facteurs qui agissent sur la saillance, on peut faire décroître la saillance perceptive proportionnellement au taux de similarité. La figure 3.3 présente des exemples de facteurs préattentifs comme la netteté et la couleur permettant d'organiser les réponses par saillance perceptive décroissante.



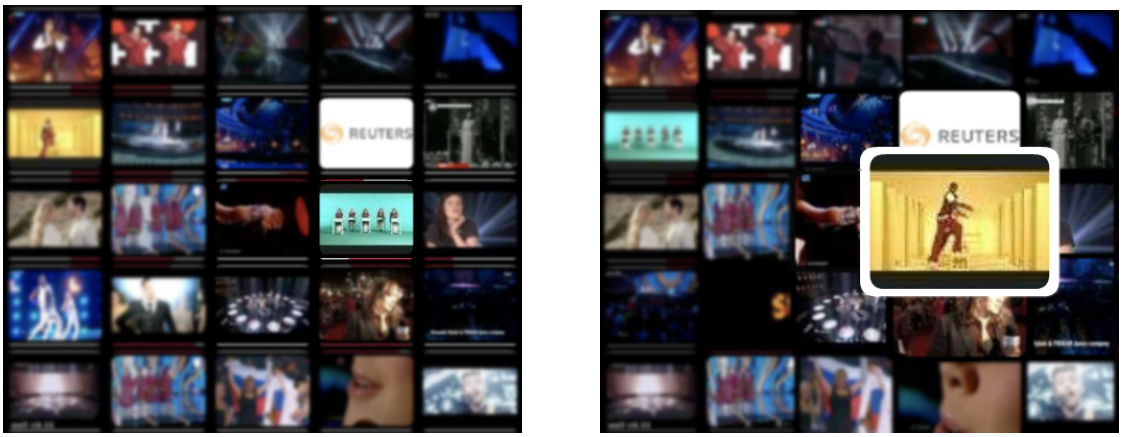

**(a)** Netteté. **(b)** Saturation.

**Figure 3.3.** Hiérarchisation d'éléments par saillance perceptive appliquée à l'affichage de documents multimédia. Ici la requête "Eurovision" sur Youtube.

L'exemple choisi reprenant l'affichage page par page des résultats selon l'affichage "*Youtube*", il y a évidemment redondance entre l'indice (couleur ou netteté) et la position verticale du document. Cela permet cependant d'illustrer plus facilement le principe mis en jeu qui peut bien évidemment être appliqué sur une présentation simultanée de vidéo telle que le mur de *BlinkX*. On aura alors la possibilité de fusionner l'effet de hiérarchisation avec les outils de

grossissement sous le curseur (voir figure 3.4).

Ce type d'affichage pourrait être un avantage dans le cas où l'utilisateur n'a pas de connaissance préalable de ce qu'il recherche (*browsing*) : cet outil permettrait de faciliter la séparation entre ce que l'utilisateur choisit d'analyser (le focus) de ce qu'il rejette temporairement (le contexte). Dans ce cas, on peut supposer que rendre saillante la zone située sous le curseur (ou avec un détecteur de mouvement oculaire la zone située dans la direction directe du regard) faciliterait la recherche.




**(a)** Sans grossissement. **(b)** Combinaison avec le grossissement.

**Figure 3.4.** Augmentation de la saillance par la netteté sous le curseur avec une présentation de documents vidéos comme celle du mur de BlinkX.

Nous nous sommes alors interrogée sur la validité de cette hypothèse et nous avons évalué le rôle d'un paramètre préattentif comme facilitateur de recherche visuelle dans un cas où tous les distracteurs et les cibles sont complexes (comme c'est le cas avec des données multimédia). Cette question sera l'objet d'étude du chapitre 6 où nous présenterons l'évalutation du flou comme facilitateur de recherche d'images fixes.

Enfin, nous avons vu, dans la section précédente, que différents paramètres visuels pouvaient favoriser la détection d'une cible présentée au milieu de plusieurs distracteurs visuels. Dans le cadre de nos travaux sur une recherche multimédia et non plus visuelle seule, nous nous sommes demandée s'il existe en audio des traits caractéristiques sonores qui fassent « sauter aux oreilles » un ou plusieurs sons parmi un ensemble de sons simultanés et que l'on pourrait combiner aux traits caractéristiques visuels. C'est le sujet de la prochaine section.

## 3.3 Principes perceptifs lors de l'écoute de plusieurs sons simultanés

Au début de ce chapitre, nous avons vu différents paramètres qui influencent la recherche d'un objet visuel présenté simultanément à d'autres. Différentes questions se posent alors sur la perception auditive avant de valider notre hypothèse selon laquelle on peut effectivement utiliser la simultanéité dans les interfaces audio.

1) Comment fait-on pour distinguer les différentes sources sonores présentes dans un environ-

nement complexe ?

2) Peut-on se concentrer sur une seule source quand il y a de nombreuses sources sonores autour ? Quels facteurs favorisent cette capacité ?
3) Existe-t-il un effet de *pop-out auditif* pour lequel certaines sources seraient tellement saillantes qu'elles « sauteraient aux oreilles » ?

### 3.3.1 Bases de l'analyse de scènes auditives

Lorsque nous sommes entourés de signaux sonores provenant de différentes sources, comme ce serait le cas avec une interface auditive ou audiovisuelle utilisant une présentation audio simultanée, nous ne percevons qu'un mélange de différentes informations complexes que le système perceptif doit déméler afin de déterminer la présence, la position et la nature des sources sonores présentes. Le système perceptif gère alors la séparation et l'organisation des différentes sources selon différents principes dans une étape de très bas niveau. La séparation de sources distinctes est appelée ***ségrégation*** (ou fission), tandis que la fusion de plusieurs attributs acoustiques est appelé ***groupement*** et permet la formation d'un ***flux auditif***. Tous ces principes sont regroupés sous le terme d'***analyse de scènes auditives*** (ou *Auditory Scene Analysis*) [Bregman 1990]. Les éléments sonores perçus simultanément sont regroupés selon différents paramètres structuraux : ceux qui ont une cohérence structurale forment une unité psychologique permettant ainsi la perception d'une source sonore distincte. Ces paramètres structuraux sont principalement :

- le **contenu fréquentiel**, notamment la hauteur spectrale (*pitch*), l'enveloppe spectrale et l'harmonicité ;
- l'**écart et l'évolution temporels**, à savoir l'attaque (*onset*), la chute (*offset*), et les modulations d'amplitude : ainsi des éléments qui sont émis en même temps mais n'évoluent pas de manière parallèle seront perçus comme distincts ;
- la **séparation spatiale** : des éléments qui sont émis de la même position seront regroupés.

Si selon Bregman, le groupement et la ségrégation se font de manière automatique et involontaire, nous noterons toutefois que ces processus peuvent être modulés par l'attention volontaire [Carlyon *et al.* 2001].

### 3.3.2 Choisir quelle source écouter : l'attention sélective

#### 3.3.2.1 L'effet cocktail party

Plusieurs études ont montré la capacité du système auditif, dans un environnement bruyant, à se concentrer sur une seule source sonore et à ignorer [4] le reste des sources présentes. C'est un phénomène d'***attention sélective*** souligné en 1953 par Cherry sous le nom d'***effet cocktail party*** [Cherry 1953; Arons 1992; Bronkhorst 2000]. En effet, lors de réceptions (*cocktail party*), nous sommes capables, sous certaines conditions, de n'écouter que notre interlocuteur alors même que de nombreuses conversations ont lieu autour, et que le bruit de fond de la salle est important.

4. La mise à l'écart par le système perceptif d'évènements (stimuli ou attributs) ou de réponses non pertinent(e)s pour la réalisation d'une tâche donnée est appelée ***inhibition***.

L'effet cocktail party a d'abord été mis en avant par un paradigme d'***écoute dichotique*** [Cherry 1953]. Le sujet porte un casque de sorte que chaque oreille reçoive un message sonore différent. On lui demande de se concentrer sur l'une ou l'autre oreille (attention sélective) ou sur les deux oreilles à la fois (attention divisée). Dans le cas de l'attention sélective, l'oreille sur laquelle l'attention est portée est appelée « oreille attentive », tandis que l'autre oreille est appelée « oreille ignorée ». L'exemple audio 3.1 présente un exemple de condition dichotique en anglais où chaque oreille reçoit un message différent. L'expérience menée par Cherry révèle alors que :

- le message est mieux perçu si le sujet a focalisé son attention (attention sélective) que s'il a cherché à comprendre les deux conversations (attention divisée) ;
- il est quasiment impossible de se rappeler le contenu sémantique du message dans l'oreille ignorée. Par exemple des sujets anglais ne notent même pas que le locuteur de l'oreille ignorée parle en allemand ou que le signal est lu à l'envers dans cette oreille ;
- quelques aspects primaires du signal de l'oreille ignorée restent cependant accessibles, ainsi on remarque un changement dans le genre du locuteur (passage d'une voix de femme à une voix d'homme ou l'inverse).

Par la suite, d'autres études ont approfondi par généralisation cette notion d'attention sélective. Elle désigne le processus par lequel un auditeur va allouer plus de ressources attentionnelles au traitement des sources sonores dont les caractéristiques correspondent à celles de la source sur laquelle il veut se concentrer. Cela implique que l'auditeur sache quelle source il veut écouter et qu'il en connaisse les caractéristiques acoustiques. Dans l'expérience de Cherry, il s'agissait de porter son attention sur une oreille, mais de façon plus générale l'auditeur peut diriger son attention sur une zone de l'espace [Spence et Driver 1994]. Lorsque les différentes sources sont spatialisées, l'auditeur peut aussi projeter son attention sur une zone fréquentielle [Woods *et al.* 2001; Cusack *et al.* 2004] ou sur certains timbres. C'est ce qui se passe, par exemple, lors de l'écoute d'une pièce de musique spatialisée où l'auditeur cherche à se concentrer sur un certain instrument [5]. Les sources sonores apparaissant dans la zone de focus sont traitées plus rapidement que les autres.

L'étude de la perception de la parole en milieu bruité et de la superposition de flux audio de même type (plusieurs conversations simultanées) est à l'origine de toutes les interfaces sonores basées sur la simultanéité des sons (chapitre 2, page 19). Dans notre cas, si un individu est capable de se concentrer sur une seule conversation parmi plusieurs, nous pouvons supposer qu'il sera capable de se concentrer sur le son d'une vidéo si celle-ci est présentée simultanément à d'autres vidéos contenant elles aussi un flux audio. C'est cette capacité auditive que nous cherchons donc à exploiter pour créer une interface où les documents audiovisuels sont présentés simultanément. Toutefois, les différents résultats montrent que si l'on demande à nos utilisateurs de se concentrer volontairement sur le contenu audio d'une seule vidéo à la fois, le contenu audio des autres vidéos ne leur sera pas, perceptivement, totalement accessible au même moment.

Cet effet est une confirmation qu'il est possible pour un utilisateur de n'écouter qu'une seule

5. Cette séparation par timbre est beaucoup plus complexe si la pièce de musique est diffusée par un système monophonique car les timbres des différents instruments ont tendance à se fondre pour former un timbre « orchestral ».

vidéo à la fois et ainsi d'ignorer le reste, même si l'environnement est bruyant. Il nous reste maintenant à étudier quelles sont les limites de cette capacité du système auditif et les facteurs qui peuvent rendre la perception d'un flux audio dans un environnement bruité moins coûteuse en ressources attentionnelles.

#### 3.3.2.2 Facteurs influents et limites de l'attention sélective en audition

Bien qu'on arrive à se concentrer sur une discussion particulière, nous continuons, dans une certaine mesure, de percevoir et de traiter les sons environnants. L'attention sélective n'applique donc qu'un filtrage partiel sur l'information entrante. Notre attention peut alors être captée par une autre conversation (par exemple quand on entend notre prénom, [Wood et Cowan 1995]) ou par un son inattendu (une porte qui s'ouvre, par exemple). Cette attraction vers un autre son malgré notre focus attentionnel dépendra du ***niveau d'alerte*** du nouveau son. Le niveau d'alerte peut être absolu, lorsque le signal est intense ou très différent des informations sur lesquelles se portaient notre attention ; ou il peut être relatif, c'est notre culture et notre apprentissage qui donnent au signal son importance (exemple du signal familier comme le prénom). Ces travaux montrent, d'une part, qu'il est difficile de focaliser volontairement et en permanence notre attention sur un seul flux sonore. D'autre part, ils sous-entendent qu'on peut établir à l'avance des paramètres qui augmentent le niveau d'alerte d'un son et permettent de diriger involontairement l'attention sur une nouvelle source audio plus saillante comme les traits caractéristiques visuels nous attirent vers une cible visuelle saillante (voir section 3.1).

Par ailleurs, changer volontairement notre focus attentionnel, c'est-à-dire choisir d'écouter un autre flux sonore quand on pense que l'information importante se situe finalement ailleurs, ne se fait pas immédiatement mais seulement après une certaine inertie, d'un ordre de grandeur d'environ 100 ms d'après [Koch *et al.* 2011].

Une autre des limites de l'attention sélective en audio est que plus le nombre de conversations simultanées augmente, plus il est difficile de se concentrer sur un seul flux.

La capacité du système perceptif à ignorer les sons non désirés dépend aussi de facteurs physiques du système auditif, ainsi les mal-entendants et les personnes âgées auront plus de difficultés à focaliser leur attention auditive.

A l'opposé, les capacités auditives sont nettement améliorées dans une condition cocktail party si les sources sonores sont séparées spatialement. C'est le cas le plus commun dans un environnement réel, mais ce n'est pas toujours respecté lors de la diffusion sonore dans les interfaces auditives puisque la lecture monophonique est souvent utilisée. Si l'on reprend les deux messages de l'exemple audio 3.1 et qu'on les superpose à la même position centrale dans une écoute monophonique (exemple audio 3.2), la séparation des deux messages et l'écoute sélective d'un seul des deux messages sont clairement plus compliquées. La séparation spatiale des sources est donc essentielle pour augmenter les capacités d'attention sélective de l'utilisateur. 3.1 3.2

De plus, choisir d'écouter une source sonore donnée et de rester focalisé sur elle (attention sélective volontaire) est plus facile si les sources sonores environnantes ne risquent pas de venir perturber la focalisation. Il suffit que les autres sources sonores, alors de moindre intérêt pour l'auditeur, aient une faible capacité à attirer involontairement l'attention, c'est-à-dire que leur niveau d'alerte soit faible, ou formulé autrement, que leur saillance soit faible devant la saillance

de la source sonore cible.

### 3.3.3 Saillance auditive et recherche sonore

On a vu précédemment (section 3.1) que plusieurs paramètres visuels peuvent augmenter la saillance visuelle de certains éléments d'une scène complexe de façon automatique, sans que l'observateur ait besoin de se concentrer. Plusieurs études sur la perception sonore ont cherché à montrer que de tels paramètres existent aussi dans le domaine audio. Notamment, les paramètres qui favorisent la ségrégation sonore augmentent aussi la saillance du son.

Un des paramètres les plus intuitifs pour renforcer la saillance auditive est le volume sonore. Cet effet apparaît lorsque la source cible est suffisamment plus forte que les distracteurs, c'est-à-dire quand le rapport cible-sur-distracteurs (*target-to-masker ratio*) est grand [Brungart et Simpson 2005]. Ce principe est à la base des applications ***SonicBrowser*** et ***SoundTorch*** décrites chapitre 2, section 2.2.

D'autres traits caractéristiques auditifs peuvent également donner un effet de « *pop-out* » en audio. D'après une étude de Cusack et Carlyon [Cusack et Carlyon 2003], des tons modulés temporellement émergent d'un fond sonore constitué de sons stationnaires. De même, des sons de durée longue ressortent parmi les sons courts (et inversement). Cette étude porte surtout sur la notion d'asymétrie telle qu'elle apparaît aussi en perception visuelle (voir section 3.1.2, page 33) : ainsi il est plus facile de repérer un son long parmi des sons courts que l'inverse et un son modulé parmi des sons stationnaires que l'inverse.

Cependant, du fait de l'aspect temporel des sons, les études en attention auditive portent souvent sur l'émergence d'un son à un moment donné, c'est-à-dire dans une présentation séquentielle des sons, et non sur la saillance d'un son dans une scène sonore complexe ou un mélange de sons simultanés. Au contraire, dans cette thèse et notamment au chapitre 6, nous chercherons à mettre en avant et à exploiter des traits caractéristiques auditifs provoquant un effet de *pop-out* dans une présentation simultanée de sources sonores. Nous pourrons alors réaliser des interfaces de recherche d'un document audiovisuel saillant parmi plusieurs documents audiovisuels présentés simultanément.

### 3.3.4 Perception sonore de l'espace

Comme nous l'avons vu dans le chapitre 2, les stratégies de présentation auditive reposent pour la plupart sur une présentation spatialisée des données sonores. De plus nous venons de voir dans ce chapitre sur la perception que la séparation spatiale améliore non seulement notre capacité à distinguer les différentes sources sonores les unes des autres mais facilite l'attention sélective sur une source parmi plusieurs. Comme c'est aussi un facteur qui améliore la recherche d'objets sonores [Eramudugolla *et al.* 2008], nous présentons ici les bases de la perception auditive de l'espace à travers les indices qui nous permettent de percevoir la localisation des différentes sources. Nous présentons ensuite succinctement les méthodes de restitution sonore permettant de spatialiser virtuellement les sources.

#### 3.3.4.1 Indices de la localisation sonore

Un être humain est capable, dans une certaine mesure, de déterminer la position d'une source sonore. Nous présentons ici les différents indices acoustiques, binauraux et spectraux, utiles à la localisation auditive.

##### a) Perception de la direction en azimut

Pour déterminer la direction de provenance d'une source sonore dans le plan horizontal, c'est-à-dire l'***azimut***, le système auditif compare les signaux sonores arrivant à chaque oreille. Les différences entre ces signaux « gauche » et « droite » sont appelés ***indices binauraux*** ou ***différences interaurales***. On distingue alors :

**Les différences interaurales de temps** (ou ITD pour *Interaural Time Difference*) : le temps de parcours des ondes sonores de la source à chacune des deux oreilles étant différent, le moment d'arrivée du signal sonore diffère d'une oreille à l'autre : plus la source est située à droite de l'auditeur et plus l'oreille droite de l'auditeur perçoit le signal en avance par rapport à l'oreille gauche. La figure 3.5 présente une schématisation simplifiée de cette différence de parcours.

**Les différences interaurales d'intensité** (ou ILD pour *Interaural Level Difference*) : la tête de l'auditeur cause une atténuation des ondes sonores, surtout pour les ondes de fréquences supérieures à 1.5 kHz (longueur d'onde inférieure aux dimensions de la tête de l'auditeur). L'onde sonore qui arrive à l'oreille du côté opposé à la source subit une plus grande atténuation. Ainsi, pour reprendre l'exemple précédent, si la source est située à droite de l'auditeur, le niveau sonore de l'onde arrivant à l'oreille gauche sera moins élevé que celui de l'onde arrivant à l'oreille droite.

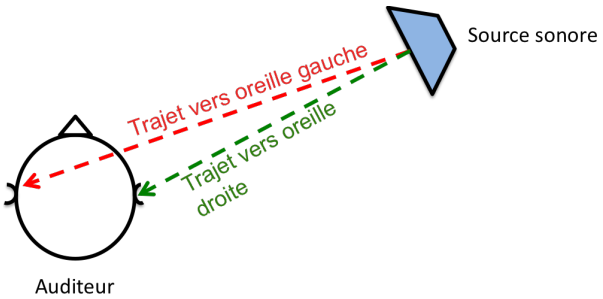



**Figure 3.5.** Schéma des parcours d'une onde sonore partant de la source S vers chacune des deux oreilles si l'on ne tient pas compte de la diffraction par la tête.

Seulement, il existe une infinité de directions pour lesquelles l'ITD et l'ILD sont constants [6]. L'ensemble de ces directions forme une surface « imaginaire » qui a la forme d'un cône ayant pour axe l'axe interaural et pour sommet le conduit auditif d'une des deux oreilles. C'est ce qu'on appelle un « cône de confusion ». Ces cônes de confusion sont responsables notamment des « confusions avant-arrière » dans le plan horizontal, et expliquent en partie les erreurs de localisation en élévation.

6. Il s'agit ici d'une approximation qui modélise la tête de l'auditeur par une sphère et ne prend pas en compte l'asymétrie de la tête.

Pour lever l'ambiguïté des indices binauraux, le système auditif dispose aussi d'indices spectraux. Ces indices sont décrits, pour chaque direction, par un couple de fonctions de transfert (une fonction par oreille) appelées ***HRTF*** (*Head Related Transfer Functions*). Ces fonctions regroupent tous les modificateurs, qui diffèrent pour chaque position de source, entre la source et nos tympans (diffractions sur notre torse et tête, masquage par la tête, filtrage fréquentiel par les pavillons, etc...) et rendent ainsi compte du filtrage obtenu en réalité entre une source et nos tympans. Comme ces modifications sont dues à la morphologie de l'auditeur, ces HRTF sont individuelles.

C'est en combinant l'utilisation des indices spectraux (HRTF) et binauraux (ITD, ILD) que le système auditif humain arrive à estimer la direction d'une source sonore dans le plan horizontal. La précision de localisation est mesurable par le *Minimum Audible Angle* tel que défini dans [Mills 1972]. Il s'agit de la plus petite différence perceptible par l'auditeur (ou *jnd* pour *just noticeable difference*). Cette précision dépend du type de stimulus et de l'angle de départ. En ce qui concerne la localisation dans le plan horizontal, un être humain est ainsi plus précis pour localiser une source frontale ($jnd \sim 3°$) que latérale ($jnd \sim 10°$).

#### b) Perception de la direction en élévation

On appelle plan médian le plan vertical qui sépare la tête en deux. Les ondes issues d'une source sonore située dans ce plan arrivent à chacune des deux oreilles de l'auditeur en même temps et avec la même intensité [7]. Il n'est alors plus possible d'utiliser les indices binauraux pour estimer l'élévation d'une source sonore. Seuls les indices spectraux sont disponibles. L'estimation de l'élévation est alors moins précise que celle de l'azimut [Begault 1994]. Elle dépend en plus de la source sonore, c'est-à-dire de sa sémantique et de son contenu acoustique comme le contenu spectral, et de la connaissance préalable du contenu spectral de cette source. Dans le plan médian, la précision est de 4° pour un bruit blanc mais seulement de 17° pour une voix inconnue [Lokki et Gröhn 2005; Rébillat 2011, p. 178].

#### c) Perception de la distance

Les différents indices acoustiques pris en compte par le système auditif sont présentés notamment dans la revue de littérature [Zahorik *et al.* 2005]. Nous retiendrons :

- **le niveau sonore (ou intensité) :** Le niveau sonore décroît avec la distance (chaque doublement de distance implique une perte de 6 dB). Cet indice permet ainsi une estimation fine de la distance entre deux sources (estimation relative « plus proche / plus lointaine ») ou les variations de distance d'une source. En revanche, sans connaissance préalable de la source, cet indice ne permet pas une estimation absolue de la distance.
- **le rapport champ direct sur champ réverbéré :** Lorsqu'il arrive aux oreilles de l'auditeur, le son d'une source située dans un milieu réverbérant est constitué à la fois du champ direct, ensemble des ondes émises par la source et arrivant directement aux oreilles de l'auditeur, et du champ réverbéré, ensemble des ondes issues de la source mais réfléchies sur les parois avant leur arrivée. Plus la source est éloignée et plus l'énergie du champ réverbéré croît par rapport à l'énergie du champ direct. Le rapport d'énergie entre les

7. Le plan médian peut en effet être vu comme un cas de cône de confusion extrême

deux champs permet donc de déterminer l'éloignement de la source dans une salle. On peut également influencer la perception de distance en ajoutant plus de réverbération à un signal sonore.

- **la composition spectrale du signal et sa familiarité :** Les hautes fréquences sont atténuées plus rapidement que les basses fréquences par la propagation dans l'air. Ainsi un son riche en hautes fréquences va renforcer la sensation de proximité. De plus lorsque les sources sont très proches, certaines distorsions spectrales qui renforçent les basses fréquences peuvent avoir lieu (effet d'emphase), perturbant notre estimation de la distance. Ainsi la composition spectrale du signal va influencer notre perception de la distance. Cependant pour être utilisée comme indice acoustique de la distance, la composition spectrale est dépendante de la connaissance préalable de la source dont il faut connaître le timbre. Par exemple, si une personne utilise un même niveau sonore pour parler normalement et pour chuchoter, elle paraîtra plus proche de l'auditeur en chuchotant. Cette mise à proximité est d'ailleurs couramment utilisée au théâtre pour prendre à partie le public.

- **les indices binauraux :** Lorsque la source est loin de l'auditeur (au-delà de 1,5 m), les indices binauraux ITD et ILD indiquant une direction d'azimut sont indépendants de la distance. En revanche pour des petites distances (en-deçà de 1,5 m), les ITD et ILD d'une même direction varient avec la distance. Ainsi les indices binauraux peuvent être utilisés pour estimer la distance en champ proche.

- **la parallaxe (indices dynamiques) :** L'effet de parallaxe désigne le fait que si un auditeur se déplace latéralement, il aura l'impression que les sources proches ont été plus déplacées que les sources lointaines. Effectuer un mouvement à gauche et/ou à droite, permet donc de comparer les déplacements relatifs entre les sources proches et les sources lointaines et ainsi d'estimer la distance d'une source sonore.

Nous avons déjà vu au chapitre 2, section 2.2 que les interfaces audio bénéficiaient d'un rendu sonore spatialisé afin de séparer au mieux les différentes sources sonores entre elles.

Toutefois, concernant la mise en avant de certaines sources sonores, c'est-à-dire l'augmentation de la saillance d'une source, seul le paramètre d'intensité a été utilisé. C'est le principe des applications de type focus+contexte ***SonicBrowser*** et ***SoundTorch*** : la source la plus forte est la plus importante. On peut littéralement parler ici de « mise en avant » puisque ce paramètre est en effet un indice d'estimation de la distance.

Nous remarquons alors que les autres paramètres, notamment contenu spectral et réverbération, n'ont été que peu utilisés à cet effet dans les interfaces sonores. Nous avons donc cherché comment il était possible d'exploiter ces paramètres pour augmenter la saillance d'une source. C'est l'objet de l'étude présentée au chapitre 6 dans laquelle nous avons évalué l'effet de mise en avant d'une source sonore contenant plus de hautes fréquences que les sources sonores concurrentes.

#### 3.3.4.2 Technologies de restitution sonore spatialisée

Il existe plusieurs méthodes de restitution qui permettent de positionner virtuellement des sources sonores dans l'espace. Chacune de ces méthodes induit une qualité de rendu spatial différente (sensation de présence, précision, réalisme, largeur de la zone de rendu, largeur de la

zone d'écoute), et est donc plus ou moins adaptée à une application de recherche de documents sonores ou audiovisuels. Alors que la majorité des stratégies de présentation sonores exposées au chapitre 2, section 2.2, reposent sur un système de diffusion externe de type réseau de deux ou plusieurs haut-parleurs, il est aussi possible d'utiliser un système de diffusion porté pour une écoute au casque. Comme les méthodes de spatialisation ne sont, en général, adaptées qu'à un système de diffusion (soit haut-parleurs, soit casque), le choix du système de diffusion a un impact sur la qualité de la spatialisation.

#### a) Restitution sur haut-parleurs

La ***stéréophonie*** est la technique la plus facile à mettre en œuvre. Dans cette technique, chaque oreille reçoit un signal provenant d'un haut-parleur gauche et un signal provenant d'un haut-parleur droit. La fusion de ces deux signaux, et l'interprétation des différences entre les signaux perçus par chaque oreille, donne lieu à une illusion perceptive dans laquelle une source virtuelle dite *fantôme* est entendue entre les deux haut-parleurs. L'inconvénient majeur de cette technique est que la position des enceintes et de l'auditeur est cruciale. Les deux haut-parleurs et l'auditeur doivent être placés au sommet d'un triangle équilatéral pour que la distance entre l'auditeur et chacun des haut-parleurs soit identique et que la synthèse de l'espace sonore soit correcte. La zone d'écoute, appelée *sweet-spot*, où l'auditeur peut se placer est donc extrêmement limitée.

La stéréophonie est également généralisable à un réseau de haut-parleurs 2D (exemples : systèmes 5.1, 7.1, *etc*) voire 3D comme c'est le cas de la technique ***Vector Based Amplitude Panning*** (VBAP) [Pulkki 1997]. Alors qu'en stéréophonie, une paire d'enceintes permet de restituer des sources frontalement et dans le plan horizontal, avec le VBAP, cette paire de haut-parleurs est remplacée par un triplet, ce qui permet de restituer des sources dans une partie plus large de l'espace et en particulier en élévation. En associant plusieurs de ces triplets, on parvient même à avoir une spatialisation en 3D tout autour de l'auditeur (périphonie). Comme en stéréophonie, la reproduction est limitée à une zone restreinte. En revanche, comme toutes les autres méthodes de stéréophonie généralisée, elle présente l'avantage de nécessiter peu de puissance de calcul et est facile à mettre en œuvre. Nous n'avons cependant pas retenu cette méthode car elle nécessite un espace avec plusieurs haut-parleurs de positions connues et semble donc peu adaptée pour une utilisation « grand public ».

L'***holophonie*** permet de s'affranchir en grande partie du problème de *sweet spot* en reproduisant physiquement le champ sonore dans une grande zone de l'espace. La ***synthèse ambisonique*** et la ***synthèse de champ sonore*** (*Wave Field Synthesis* ou WFS) sont les deux catégories de techniques holophoniques.

Le système Ambisonic fut développé par Michael Gerzon [Gerzon 1985] au départ pour de la prise de son spatialisée et approfondi plus tard pour de la ***synthèse ambisonique***. C'est la première approche à chercher à reproduire le champ acoustique plutôt qu'à spatialiser de façon indépendante différentes sources sonores. L'approche ambisonique repose sur deux phases indépendantes. La première phase, dite d'***encodage***, est indépendante du système de restitution et repose uniquement sur la position des sources sonores. Elle est basée sur une décomposition

en harmoniques sphériques. Pour augmenter la résolution spatiale de cette décomposition, il faut utiliser des systèmes d'ordre supérieur qui nécessitent plus de calculs [Daniel 2000]. La seconde phase est une phase de ***décodage***. Elle repose cette fois sur le nombre et la position des haut-parleurs pour la diffusion. L'utilisation de cette technique est relativement simple et permet de reproduire le champ sonore avec n'importe quel nombre de haut-parleurs (au moins 2) et n'importe quelle configuration, ce qui rend le système de diffusion modulable. Toutefois une répartition des haut-parleurs sur une sphère est conseillée (équidistance des haut-parleurs par rapport à l'auditeur). De plus, il est à noter qu'une augmentation du nombre de haut-parleurs permet un rendu spatial plus précis. Un des gros avantages de la synthèse ambisonique est que le temps de calcul n'augmente pas lorsque le nombre de sources augmente. En effet, la partie encodage, très faible en coût de calcul, est alors *quasi* indépendante du nombre de sources. Quant à la partie décodage, elle ne dépend que du nombre de haut-parleurs. Ainsi la synthèse ambisonique est une technique qui convient très bien à un projet contenant beaucoup de sources sonores comme c'est le cas de l'environnement que nous avons construit dans notre étude sur l'exploration de collections multimédia au chapitre 4.

Pour des ordres supérieurs suffisamment élevés, la synthèse ambisonique peut atteindre une précision spatiale équivalente à celle de l'autre technique holophonique, la ***WFS***. Cette dernière est basée sur le principe de Huygens-Fresnel : si l'on souhaite reproduire l'action acoustique d'une source sur un volume de l'espace, il suffit juste de savoir reproduire au niveau de la surface délimitant le volume étudié les ondes émise par la source [8] [Berkhout *et al.* 1993]. Donc le contrôle d'émetteurs acoustiques sur une surface délimitant un volume permet de reproduire un champ sonore donné au sein de ce volume. Comme cette technique est plus complexe à mettre en œuvre que les précédentes, on limite en pratique, pour l'instant, la restitution des sources au seul plan horizontal. L'auditeur dispose alors d'une réelle perspective sonore au sein de laquelle il peut se déplacer librement car ce système de reproduction propose une zone d'écoute bien plus large que les autres techniques de reproduction. Cette zone correspond en fait au volume dans lequel le rendu est considéré comme « exact ». D'autre part la résolution spatiale des sources offertes par cette technique est très fine. Malheureusement, la WFS a un coût élevé tant d'un point de vue matériel (nombre de haut-parleurs nécessaires + système de calcul dédié) qu'en terme de temps de calcul. Elle est donc très difficile à mettre en œuvre.

#### b) Restitution sur casque

Motivée par le développement d'outils destinés à l'exploration de collections multimédia, et en particulier pour le grand public, nous nous sommes concentrée sur des méthodes de restitution sur casque où l'équipement est plus restreint que pour les méthodes de spatialisation sur haut-parleurs. En effet, de nombreux utilisateurs possèdent aujourd'hui des ordinateurs personnels portables dont le système d'écoute principal est géralement constitué de deux haut-parleurs uniquement (ce qui limite la restitution à la stéréophonie [9]) situés très proches l'un de l'autre

8. Le principe de Huygens-Fresnel peut s'énoncer par : « Chaque point M recevant une onde issue d'une source primaire se comporte comme une source secondaire rayonnant une onde sphérique de même fréquence, même amplitude et même phase que l'onde reçue. Les fronts d'onde créés par les sources secondaires interfèrent entre eux pour recréer un front d'onde identique à celui émis par la source primaire ».

9. Nous ne tenons pas compte ici des techniques dites transaurales qui permettent de restituer du binaural sur des enceintes (par annulation des trajets croisés) mais qui sont encore peu fiables et très délicates à mettre

ce qui limite le sweet-spot. Et même dans le cas où l'utilisateur est équipé d'enceintes externes, il est rare qu'il en possède plus de deux et qu'elles soient bien positionnées.

La ***stéréophonie sur casque*** est la technique la plus facile à mettre en place puisqu'elle reprend les principes de la stéréophonie sur haut-parleurs sans adaptation particulière. Cependant, l'isolation de chaque canal gauche/droite produit alors un problème d'externalisation : les sources sonores sont perçues à l'intérieur de la tête.

Le ***rendu binaural*** résout, au moins partiellement, le problème d'externalisation [Begault 1994; Nicol 2010]. La synthèse binaurale permet de construire virtuellement un espace sonore à partir des informations morphologiques de l'utilisateur. Il s'agit de reproduire le champ sonore tel qu'il serait perçu à l'entrée des oreilles de l'auditeur. Le rendu sonore est basé sur un filtrage du signal monophonique à spatialiser par un couple de HRTF. Comme les HRTF dépendent de la morphologie de chaque individu, idéalement le traitement devrait être adapté à l'utilisateur. De nombreux travaux se penchent sur ces problèmes d'individualité des HRTF afin de rendre possible l'utilisation d'HRTF mesurées sur un autre individu voire sur un mannequin. Des solutions mathématiques ont été proposées ainsi que des techniques d'apprentissage [Parseihian et Katz 2012b]. Nous ne nous sommes pas penchée sur ces questions de l'individualisation car nous avons jugé les sensations d'espace et de séparation de sources suffisantes pour l'utilisation que nous avions à en faire, même avec un rendu binaural non indivudualisé. De nombreux exemples d'enregistrement ou de synthèse binaurale sont disponibles sur Internet, comme cet enregistrement de son de boite d'allumettes qui tourne virtuellement autour de l'auditeur (écouter l'exemple audio 3.3 [10]). De plus, lorsque l'utilisateur est amené à tourner la tête ou à se déplacer dans un environnement virtuel, un suivi des mouvements de la tête est nécessaire (*tracking*). Utilisant un rendu visuel de petite taille (écran d'ordinateur personnel), nous avons jugé que les mouvements de tête étaient relativement restreints et que l'absence de tracking ne nuirait pas aux techniques proposées dans la suite du manuscrit. Le principal défaut de cette méthode binaurale réside surtout dans les ressources nécessaires pour un calcul en temps réel qui sont assez importantes et qui augmentent avec le nombre de sources (le filtrage dépend de la position de chaque source). Pour pallier ce problème de coût de calcul, nous avons utilisé, lorsque nous avions besoin de spatialiser de nombreuses sources (par exemple 100 sources simultanées au chapitre 4), une méthode d'Ambisonic virtuel.

3.3

La séparation entre l'encodage et le décodage dans la synthèse ambisonique permet de remplacer la phase de décodage ambisonique (pour un rendu sur haut-parleurs) par un décodage binaural pour une restitution au casque. On parle d'***Ambisonic virtuel*** [Noisternig *et al.* 2003]. Cela revient à spatialiser en binaural, non plus chaque source sonore à une position différente, mais ce qu'aurait émis chaque haut-parleur dans une restitution ambisonique. On peut ainsi combiner les avantages de la synthèse ambisonique et du binaural : le temps de calcul reste faible même pour un très grand nombre de sources (encodage ambisonique) mais la restitution se fait sur casque et limite donc l'équipement nécessaire (décodage binaural).

---

en place.

10. extrait de http://www.youtube.com/watch?v=x5G3HUiscW4

## 3.4 Perception multisensorielle

Dans notre vie courante, les informations que nous recevons du monde extérieur nous parviennent par différents canaux sensoriels (comme la vision ou l'audition) que nous appellons modalités sensorielles ou plus simplement modalités. Notre système perceptif doit alors traiter ces différentes informations et garantir une perception robuste, cohérente et ainsi permettre une meilleure interaction avec l'environnement.

Dans certains cas, les informations seront combinées voire fusionnées (*binding*) : c'est ce qu'on appelle l'intégration multisensorielle. Nous percevons alors un seul objet audiovisuel au lieu d'un objet auditif d'une part et d'un objet visuel distinct d'autre part. Dans d'autres cas, les informations ne seront pas liées mais la perception d'une modalité sera influencée par la présence d'une autre modalité même non pertinente : on dit qu'il y a interaction entre les modalités. Ernst et Bülthoff [Ernst et Bülthoff 2004] distinguent ces deux processus perceptifs d'après la redondance ou non entre les informations issues des deux modalités : l'*intégration* sensorielle est une interaction entre signaux redondants tandis que l'interaction entre signaux non redondants est appelée *combinaison*.

Ces interactions sensorielles provoquent une réponse de notre système perceptif différente de la réponse sensorielle produite dans un cas unimodal. Concernée par la recherche de documents vidéos contenant uniquement des informations visuelles et sonores, nous illustrerons ici nos propos par des exemples en rapport à l'intégration audiovisuelle. Toutefois nous noterons que des interactions entre les autres sens humains existent aussi [Calvert *et al.* 2004; Stein et Meredith 1993], par exemple des interactions entre sens tactile et visuel [Shams et Kim 2010] ou entre sens tactile et auditif [Occelli *et al.* 2011; Kitagawa et Spence 2006].

Concernant les interactions audiovisuelles, le lecteur pourra également se référer aux revues de littérature de [Spence 2007 2011] et [Ghirardelli et Scharine 2009].

### 3.4.1 Impacts de la multimodalité

Les phénomènes d'intégration multisensorielle et les influences d'une modalité sur l'autre ont souvent été mis en avant par des expériences où les deux modalités étaient présentées dans un rapport conflictuel, chacune présentant une information différente. De ces situations conflictuelles découlent alors des illusions perceptives.

#### 3.4.1.1 Illusions perceptives

##### a) Effet McGurk

La présentation de stimuli visuels et auditifs non cohérents peut produire un tout autre percept ne correspondant ni au stimulus visuel ni au stimulus sonore. C'est ce que l'on observe avec l'***effet McGurk*** [McGurk et MacDonald 1976] où la partie visuelle d'une vidéo d'une personne prononçant le phonème /*ga*/ est présentée en même temps que le son d'une personne prononçant un autre phonème, /*ba*/ : la syllabe finalement perçue étant /*da*/. Ainsi la perception auditive de la syllabe est modifiée par la perception simultanée du mouvement des lèvres. Dans des conditions normales, les informations auditives et visuelles de la parole sont complémentaires

et cohérentes. Or ici, on mélange des phonèmes qui sont visuellement ambigus (parmi les syllabes /*ga*/ et /*da*/) et des phonèmes acoustiquement ambigus (/*ba*/ et /*da*/). Le système visuel perçoit /*ga*/ ou /*da*/ tandis que le système auditif perçoit /*ba*/ ou /*da*/. Afin de satisfaire la cohérence habituelle entre les deux modalités, le système perceptif interprète le percept comme celui qui est le plus probable : /*da*/ (voir figure 3.6 et vidéo 3.1 [11]). Cela suit une règle d'optimisation de l'intégration. Ce phénomène a lieu aussi quand la syllabe visuelle /*ka*/ est présentée simultanément à la syllabe acoustique /*pa*/ : on perçoit /*ta*/. Au final, ce phénomène peut être observé également pour des phrases entières [Massaro et Stork 1998]. Ainsi la présentation de la séquence visuelle "*my gag kok me too grive*" associée à la séquence sonore "*my bap pop me poo brive*" conduit à la perception de la phrase "*my dad taught me to drive*". L'illusion est dans ce cas d'autant plus robuste que les deux séquences unimodales prises individuellement n'ont aucun sens et que seule leur fusion permet d'obtenir une phrase cohérente.

3.1

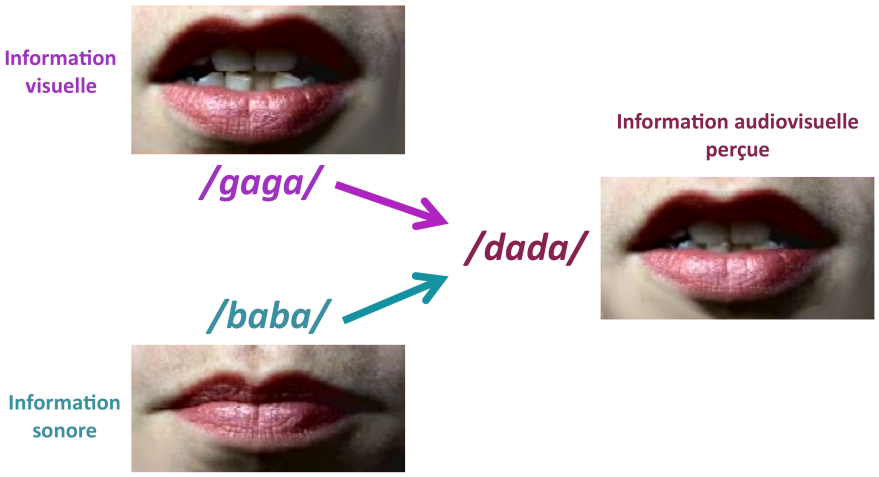



**Figure 3.6.** Illustration de l'effet McGurk : la fusion d'un stimulus visuel et d'un stimulus auditif non cohérent induit un nouveau percept.

### b) Ventriloquie

Un autre exemple très connu, l'***effet ventriloque***, révèle un biais intermodal en faveur du visuel sur une tâche de localisation. Cet effet est exploité par les marionnetistes ventriloques pour donner l'illusion que c'est leur marionnette qui parle : le ventriloque parle sans bouger les lèvres tandis qu'il fait bouger celles de sa marionnette, la parole est ainsi attribuée à cette dernière. Cet effet permet aussi de regarder un film sans être géné par la différence de localisation entre le personnage à l'écran et la position des enceintes. Plus généralement, l'effet ventriloque est valable pour toutes sortes de stimuli, pas forcément liés à la parole. Il a lieu lorsque deux stimuli visuels et auditifs sont émis avec des décalages temporels et spatiaux favorisant l'intégration multisensorielle mais qu'ils ne sont pas spatialement coïncidents. Le cas échéant, la localisation d'un son est déviée vers la position du stimulus visuel situé à proximité [Bertelson et Radeau

11. D'après la vidéo "Try the McGurk Effect : Is Seeing Believing ?" proposée par la BBC2 http://www.youtube.com/watch?v=G-lN8vWm3m0

1981; Spence et Driver 2000]. On dit qu'il y a ***capture visuelle***. Toutefois, plus l'écart temporel ou spatial entre les stimuli augmente et plus faible est le biais.

On notera que certains auteurs ont rapporté l'existence d'un biais auditif sur la localisation d'une cible visuelle [Bertelson et Radeau 1981] mais que ce biais est beaucoup plus faible que le biais visuel pour une tâche de localisation sonore. On considère qu'il y a dominance visuelle dans une tâche de localisation.

#### c) Illusions visuelles et influence de l'audition

Il existe également des phénomènes de ***capture auditive***. Par exemple, dans la perception des ***flashs illusoires*** [Shams *et al.* 2002], la présentation d'un flash lumineux unique peut être perçu comme une succession de flashs lumineux s'il est accompagné de plusieurs bips sonores. Cette illusion est illustrée à travers la vidéo 3.2 qui présente la condition standard avec un seul bip sonore et la vidéo 3.3 qui présente l'illusion obtenue avec deux bips sonores [12].

De même, l'audition peut influencer l'interprétation d'une scène visuelle ambigüe, c'est-à-dire à plusieurs interprétations plausibles. Par exemple, lors du ***bounce inducing effect*** on présente à l'écran deux sphères visuelles avançant l'une vers l'autre. Au moment où elles se rejoignent deux interprétations sont possibles : soit les sphères se croisent et continuent leur trajectoire initiale, soit elles rebondissent l'une sur l'autre et repartent dans l'autre sens. Dans un contexte unimodal visuel seul, la première interprétation est la plus commune. Pourtant quand on ajoute un son bref au moment où les deux sphères coïncident, on perçoit un rebond [Sekuler *et al.* 1997]. Là encore, plusieurs paramètres comme l'écart entre l'instant d'émission du son, la composition spectrale du son ou encore l'attention portée sur l'une ou l'autre des modalités modifient l'intensité de cet effet [Watanabe 2001].

### 3.4.1.2 Avantages de la multimodalité en situations non illusoires

Lorsque les stimuli issus des différentes modalités ne sont plus en conflit mais représentent au contraire le même évènement, on parle d'***effet de cible redondante*** (ou RSE pour *Redundant Signal Effect*). De nombreuses études, tant comportementales que neurophysiologiques, ont montré qu'un objet perçu sous deux modalités redondantes est traité plus rapidement et donne lieu à des appréciations plus justes sur son identité, sa détection ou sa position, en comparaison à un objet perçu sous une seule modalité.

#### a) Perception multimodale de la parole

L'effet McGurk, vu en section 3.4.1.1 a) , illustre bien l'influence du visuel sur la perception de la parole. Cependant il s'agit d'un phénomène illusoire non écologique, induit par une situation conflictuelle que nous ne trouverions pas au quotidien. D'autres exemples démontrent l'influence de la vision sur la compréhension de la parole dans des situations plus réelles où la vision et l'audition apportent des informations complémentaires sur le même message. En particulier il a été montré que la lecture labiale facilite le traitement de la parole dans le bruit (entre autres [Sumby et Pollack 1954; Ross *et al.* 2007; Grant et Seitz 2000]). Par exemple, Sumby et Pollack ont trouvé que la reconnaissance des indices audiovisuels de paroles se dégradent

12. La démonstration de cette illusion est issue du site http://shamslab.psych.ucla.edu/demos/, disponible à la date du 09 juillet 2012.

moins rapidement que les indices acoustiques seuls. Dans leur exemple de vocabulaire à 8 mots, l'identification acoustique seule passe d'un taux de 100 % à 0 dB à moins de 20 % à −30 dB de rapport signal-sur-bruit (RSB). En revanche en ajoutant la lecture labiale, l'identification ne baisse que de 100 % à 90 %, montrant ainsi que la présence de la composante visuelle permet un gain équivalent à environ 12 dB RSB. La vision améliore aussi la compréhension de la parole dans des conditions difficiles non liées au bruit comme c'est le cas avec des langues étrangères [Arnold et Hill 2001].

D'autres indices visuels que la lecture labiale influencent d'ailleurs l'intelligibilité de la parole, notamment la gestuelle et l'expression du visage.

#### b) Identification d'objets ou de scènes ambigus

L'avantage de la redondance entre les stimuli visuels et sonores est observable dans d'autres sitations sans rapport avec l'intelligibilité de la parole. Par exemple, la simple présence d'un stimulus sonore accessoire, c'est-à-dire ne présentant aucun lien sémantique ou structurel avec le stimulus visuel, raccourcit les temps d'identification d'un stimulus visuel, qu'il soit clairement visible [Doyle et Snowden 2001] ou masqué [Chen et Spence 2009] [13]. Cet effet de facilitation sur l'identification d'images masquées se retrouve aussi avec la présence d'un son associé sémantiquement à l'image présentée (par exemple un aboiement quand l'image représente un chien) [Chen et Spence 2010]. Miller [Miller 1982] note d'ailleurs que l'effet de facilitation est plus marqué lorsque les deux composantes audio et visuelles se rapportent au même évènement (congruence sémantique ou redondance de cibles).

Cet effet de redondance de cibles se retrouve ainsi dans d'autres conditions peu réalistes : par exemple, l'identification d'objets audiovisuels A (une ellipse verticale + un son grave) et B (une ellipse horizontale + un son aigu) prend moins de temps lorsque ces objets sont présentés à la fois en audio et en visuel [Giard et Peronnet 1999] ; ou plus réalistes : par exemple, dans une étude sur l'identification des sons d'environnement, Ozcan et van Egmond [Ozcan et van Egmond 2009] montrent une amélioration des performances, en terme de justesse et de temps de réponse, lorsqu'un contexte visuel montrant la scène dans laquelle le son est censé être émis est présent. Le gain de temps se retrouve aussi dans une étude sur la simulation réaliste d'objets en 3D [Suied *et al.* 2009]. Nous noterons d'ailleurs que, dans cette étude, la coïncidence spatiale n'a pas d'influence contrairement à la congruence sémantique. Enfin, de meilleures performances d'identification sont observées en condition bimodales qu'unimodales même si l'un des deux stimuli est présenté en amorce et non simultanément (*priming effect*) [Schneider *et al.* 2008].

#### c) Détection

La présentation bimodale améliore aussi la perception de la présence d'un stimulus, c'est-à-dire la détection. Ainsi le seuil de détection de signaux faibles diminue en présentation bimodale : des signaux que l'on aurait manqués dans une condition unimodale deviennent perceptibles en condition bimodale. L'intensité visuelle semble alors augmenter en présence d'un stimulus sonore même accessoire [Bolognini *et al.* 2005]. De même, comme on l'a vu pour la parole, la présence

13. La notion de masquage concerne ici un masquage temporel : si deux images sont présentées l'une après l'autre trop rapidement, alors la deuxième image masque perceptivement la première. L'ajout d'un son accessoire permet d'identifier la première image que le système visuel n'aurait pas perçue en l'absence de son.

d'indices visuels permettant la lecture labiale abaisse le seuil de détection de la parole dans le bruit [Grant et Seitz 2000].

#### d) Localisation

Suite aux études sur l'effet ventriloque (voir section 3.4.1.1 b) ), effet dans lequel la position perçue est déviée vers la position de la composante visuelle, de nombreuses études ont documenté la perception de l'espace en conditions multisensorielles. Les résultats montrent que l'estimation de la position audiovisuelle dépend de la fiabilité de chaque composante. Ainsi, si l'on diminue la fiabilité de la perception visuelle en appliquant du flou sur le stimulus visuel ou qu'on présente le stimulus visuel dans le champ visuel périphérique, le biais vers la position visuelle est extrêmement diminué. Burr et Alais [Burr et Alais 2006] expliquent cela par un modèle de maximum de vraisemblance (*Maximum Likelihood Estimation*, voir équations (3.1) : à partir des mesures de la position perçue pour un stimulus sonore seul (position $m_A$, variance $\sigma_A$) et des mesures de la position perçue pour le stimulus visuel associé présenté seul (position $m_V$, variance $\sigma_V$), on peut déterminer la position audiovisuelle perçue $m_{AV}$ avec une variance $\sigma_{AV}$.

$$m_{AV} = \frac{\sigma_V^2}{\sigma_V^2+\sigma_A^2} m_A + \frac{\sigma_A^2}{\sigma_V^2+\sigma_A^2} m_V \qquad \sigma_{AV}^2 = \frac{\sigma_V^2 \sigma_A^2}{\sigma_V^2+\sigma_A^2} m_A \leq min(\sigma_V, \sigma_A) \tag{3.1}$$

Le système perceptif s'appuierait donc sur l'estimation de la position visuelle $m_V$ et de sa fiabilité $\sigma_V$ ainsi que de l'estimation de la position auditive $m_A$ et de sa fiablité $\sigma_A$ pour estimer la position audiovisuelle d'un objet. Dans le cas où l'estimation visuelle est beaucoup plus fiable que la modalité auditive, $\sigma_V \gg \sigma_A$, alors $m_{AV} \simeq m_V$ et $\sigma_{AV} \simeq \sigma_V$ et la perception multimodale est dominée par la perception visuelle. En revanche en champ périphérique, la fiabilité du visuel diminue ($\sigma_V \simeq \sigma_A$) donc le système perceptif prend en compte les deux estimations à part égale. Dans tous les cas, la localisation est plus juste lorsque l'on combine une présentation visuelle et une présentation sonore [Rébillat 2011, chapitre 8, p. 205].

Toujours sur la perception de la localisation, une combinaison bimodale permet également de détecter des changements de position (stimuli déviants) plus rapidement. Dans une expérience de [Schröger et Widmann 1998], des stimuli audiovisuels constitués d'un bruit blanc et d'un flash lumineux sont répétés toujours à la même position centrale mais de temps en temps l'une ou l'autre ou les deux composantes audiovisuelles sont déviées de 15° vers la droite ou la gauche. Les participants doivent indiquer quand arrivent ces déviations. Les temps de réaction sont plus courts lorsque les deux composantes du stimulus audiovisuel sont déplacées simultanément que lorsqu'une seule des deux l'est.

### 3.4.1.3 Application de la perception multisensorielle à la recherche audiovisuelle

Nous avons vu, dans les deux premières sections de ce chapitre, les différents facteurs unimodaux qui peuvent influencer, d'une part, la recherche d'objets visuels et, d'autre part, la ségrégation de flux audio et la recherche sonore. Nous avons également présenté différentes interactions qui existent entre les modalités visuelles et sonores et leurs conséquences sur l'identification ou

la détection d'objets. La présentation qui suit a pour but d'exposer comment chacune des deux modalités audio et visuelle va influencer la recherche unimodale dans l'autre modalité.

#### a) Influence de l'audio sur une tâche de recherche visuelle

La recherche visuelle peut être améliorée par la présence de sons aux caractéristiques sonores spécifiques. On parle alors d'***aurally-aided visual search***. La plupart des études sur le sujet repose sur le rôle d'alarme et la capacité de l'audition à diriger l'attention dans une direction donnée (*spatial orienting*). Ainsi un son placé dans la direction d'une cible visuelle permet de retrouver cette cible plus rapidement [Perrott *et al.* 1991; Bolia *et al.* 1999; Simpson *et al.* 2010].

Cependant d'autres effets positifs de l'audio sur la recherche visuelle ne faisant pas intervenir la spatialisation du son ont été observés. Ainsi d'après [Iordanescu *et al.* 2010 2011], la facilitation obtenue par l'ajout d'un son sur une recherche visuelle pourrait être due à une intégration sémantique entre les composantes sonores et visuelles. Ainsi un son, non spatialisé, caractéristique d'une source sonore (exemple : le son « miaou »/mjau/) augmente la saillance d'une cible correspondante (exemple : l'image d'un chat ou le mot CHAT) et donne ainsi des temps de recherche visuelle plus courts. Dans les études de [Van der Burg *et al.* 2008], plusieurs objets visuels colorés sont présentés simultanément avec deux orientations diagonales possibles. La couleur de ces objets varie dans le temps. La tâche consiste alors à retrouver un objet ayant une orientation verticale ou horizontale et non diagonale. Ces travaux ont montré qu'en associant un stimulus auditif non spatialisé et non-informatif (ne partageant pas d'information sur la source, par exemple n'indiquant pas la couleur) on peut accélérer la recherche visuelle dans le cas où le son est synchronisé temporellement avec les changements de couleur de la cible. La cible visuelle semble alors sortir spontanément du lot et l'on peut alors qualifier cela de ***pop-out audiovisuel***.

L'ajout d'un son peut également améliorer la détection d'objets visuels en mouvement [Kim *et al.* 2011]. Lorsque plusieurs points en mouvement sont présentés simultanément avec une direction aléatoire et différente pour tous les points, sauf pour un sous-ensemble de points dits coordonnés (c'est-à-dire qu'ils se déplacent dans la même direction), l'ajout d'un son spatialisé ayant le même sens de mouvement que les points coordonnés fait ressortir ces points et rend ainsi leur détection plus facile. L'effet positif de l'ajout de son est dû, cette fois, non à une synchronisation temporelle, mais à une synchronisation spatiale.

Pour résumer, un son indiquant la direction ou le mouvement d'une cible visuelle, synchronisé avec cette cible et/ou partageant des informations sémantiques avec la cible accélèrera la recherche de celle-ci.

#### b) Influence du visuel sur une tâche de recherche sonore

Si l'ajout d'un son peut faciliter la recherche visuelle, l'inverse est aussi vrai. Notamment, les résultats d'une étude de [Best *et al.* 2007] montrent qu'ajouter un indice visuel indiquant la position et l'instant d'émission d'un son cible améliore les performances d'une tâche de recherche sonore, même si l'indice visuel n'apporte aucune information sémantique sur le son ou la source.

Cette étude repose sur une expérience où les participants ne savent ni où ni quand interviendra la cible sonore parmi plusieurs distracteurs sonores. Il s'agit d'un cas d'incertitude spatiale et temporelle : le son cible pourra être émis par l'un des 5 haut-parleurs disposés devant le

sujet et émis durant l'un des 5 segments temporels successifs possibles. Plusieurs types de sons sont alors utilisés, soit des chants d'oiseaux (cible = un chant connu, distracteurs = chants inconnus) soit de la parole (cible = nombres, distracteurs = autre parole). L'effet de facilitation visuelle est observable quel que soit le type de sons mais il est bien plus marqué pour les chants d'oiseaux moins familiers que les sons de parole. D'après les auteurs, l'indice visuel permettrait simplement de focaliser son attention au bon endroit ou au bon moment et avantagerait ainsi l'attention sélective auditive, qui permet la ségrégation sonore et donc l'identification d'une source parmi plusieurs sons concurrents.

Confirmant cet effet de facilitation dans une étude plus récente, Varghese et ses collègues [Varghese *et al.* 2012] montrent que la combinaison de l'indice visuel indiquant quand écouter et de la spatialisation sonore peut être redondante sans que cela n'amplifie le gain. Pour que l'ajout d'un indice visuel soit bénéfique, il faut que la ségrégation des sons soit difficile au départ, comme c'est le cas si les sons proviennent de la même position.

Ces effets du visuel sur l'audio et de l'audio sur le visuel démontrent la possibilité d'attirer l'attention sur une portion de l'espace grâce à l'autre modalité. On parle d'orientation spatiale intermodale (*crossmodal spatial orienting*). Cette capacité n'est pas seulement utile pour la recherche mais peut etre utile également pour des interfaces de prévention et d'alames, ou sur des postes de surveillance ou de contrôle.

### 3.4.2 Facteurs influençant l'intégration multisensorielle

Nous avons vu que l'écart spatio-temporel entre les stimuli audio et visuels pouvaient influer sur la robustesse des effets notamment d'illusions, par exemple l'effet de ventriloquie dépend de l'écart spatial. La fusion (*binding*) et l'interaction de stimuli issus de modalités sensorielles différentes n'ont lieu que sous certaines conditions et la robustesse des phénomènes observés dépend de nombreux facteurs, certains directement liés aux stimuli (processus dit *stimulus-driven*), d'autres sont cognitifs et dépendent de l'individu, de sa tâche et de ses connaissances préalables. Ces facteurs témoignent d'une certaine flexibilité dans le traitement perceptif et neuronal des informations auditives et visuelles.

#### 3.4.2.1 Facteurs structuraux (stimulus-driven)

##### a) Contiguïté spatio-temporelle

Comme pour l'intégration unimodale d'éléments visuels [Treisman 1998] ou sonores [Bregman 1990], l'intégration multisensorielle ne peut avoir lieu que sous certaines conditions de contiguïté spatio-temporelle [Lewald et Guski 2003]. Les stimuli à intégrer doivent avoir lieu à des instants proches l'un de l'autre et en des positions proches. Plus l'écart spatio-temporel est grand et moins l'intégration est robuste.

Sur le plan temporel, l'écart entre les stimulus visuel et auditif doit être compris dans des fenêtres temporelles relativement courtes. Cette fenêtre peut être plus large pour des stimuli avec un sens plus prononcé comme les indices de parole (voix+ visage/gestes). Au-delà de cet écart temporel, le décalage temporel est apparent et les signaux audio et visuels perçus semblent désynchronisés tandis qu'en-deça le décalage n'est pas perceptible [Hollier *et al.* 1999; Lewald et Guski 2003]. Les différentes études sur le sujet donnent des valeurs légèrement différentes

selon les procédures expérimentales employées ( environ [−50 ms, +100 ms] d'après [Hollier *et al.* 1999], avec pour origine de temps l'instant de présentation du stimulus visuel) avec une meilleure tolérance pour la parole puisque la fenêtre d'intégration temporelle peut alors s'étendre à [−90 ms, +180ms]. Cependant toutes les études s'accordent sur l'asymétrie de cette fenêtre puisque l'intégration se fait moins bien lorsque le stimulus auditif est en avance par rapport au stimulus visuel.

Une certaine tolérance est aussi acceptée dans l'écart spatial entre les stimuli audio et visuels à intégrer. Par exemple, lorsqu'il s'agit d'indiquer la position de la source, les composantes audio et visuelles peuvent être séparées d'environ 2 à 3° [Lewald et Guski 2003]. Au-delà on perçoit distinctement deux sources séparées une visuelle et une auditive. De plus, plus l'écart augmente et plus les effets de facilitation, sous forme de gain en temps de réponse, diminuent. Lorsqu'il ne s'agit pas de localiser la source, mais plutôt de comprendre un message, la concordance spatiale des informations unimodales semble moins essentielle pour l'intégration et l'effet de facilitation est encore observable même avec des stimuli très disparates, de l'ordre de 11° à 30° pour des stimuli avec de la parole [Soto *et al.* 2008].

#### b) Fiabilité et intensité des signaux

Certaines propriétés unimodales, indépendantes dans chaque modalité, comme l'intensité du signal ou la fiabilité de chaque signal, influencent également l'intégration multisensorielle.

Le besoin d'intégration multisensorielle baisse quand le stimulus unimodal offre une identification/réponse déjà robuste et donc non ambigüe. C'est le principe d'***efficacité inverse*** (*inverse effectiveness*) introduit par Stein et Meredith [Stein et Meredith 1993] au niveau neuronal : la réponse à un stimulus bi-, voire tri-, modal est d'autant plus forte que les différentes composantes unimodales sont peu efficaces, c'est-à-dire que l'intensité du signal est faible voire proche du seuil de détection. Si l'une des modalités est démesurément plus intense que l'autre, il y a aura donc difficilement intégration multisensorielle. Il faut donc, dans le cas où l'on ne cherche pas à obtenir de dominance de l'une ou l'autre des modalités, équilibrer l'intensité des stimuli unimodaux pour obtenir une réponse (temps de réponse, amplitude des signaux neuronaux, intensité subjective) équivalente dans les deux modalités. On parlera de ***calibration multimodale***.

De plus, lorsqu'une modalité est plus fiable que l'autre, elle prédomine dans l'intégration multisensorielle et introduit un biais. C'est ce qui se passe par exemple dans l'effet ventriloque. Par habitude (donc par apprentissage), l'être humain sait que l'estimation de la position spatiale d'une source est plus précise en vision qu'en audition. Il a donc tendance à se fier plus au visuel, et en cas de conflit, la position perçue est plus proche de la position de la source visuelle. Au contraire, toujours dans une tâche de localisation, si l'on rend flou le rendu visuel, qui devient alors moins fiable, les participants s'appuient de moins en moins sur le visuel mais au contraire plus sur l'audio [Ernst et Bülthoff 2004]. Ernst et Bülthoff ont alors introduit un modèle de perception pour expliquer ce phénomène : les signaux issus des deux modalités sont combinés par pondération selon un ***modèle de maximum de vraisemblance*** (MLE pour *Maximum Likelihood Estimation*) qui prend en compte la fiabilité de chacune des modalités. La fiabilité d'un processus est mesurée par la variance de l'estimation sensorielle. Ce mécanisme de MLE permet d'obtenir une estimation plus fiable dans la condition multisensorielle que dans chacune

des deux modalités prises une à une. Ce modèle est également cohérent avec la *Modality Precision Hypothesis*, dont Welch et Warren [Welch et Warren 1980, p. 657] exposent les grandes lignes :

> *...when two sensory modalities provide discrepant information about some characteristic of an event, the resolution of the discrepancy will favor the modality that is the more precise of the two in registering that event.*

Nous retiendrons que pour éviter d'obtenir un biais, il faut que les deux modalités conduisent à des estimations de même niveau de fiabilité.

#### 3.4.2.2 Facteurs cognitifs

##### a) Lien sémantique

L'effet de redondance de cibles présenté auparavant (section 3.4.1.2) nous permet d'affirmer que lorsque deux stimuli auditifs et visuels, coïncidants temporellement et spatiallement, sont reliés sémantiquement et représentent ainsi le même concept, on observe un gain de temps dans l'identification de l'évènement ou de la source dont sont issus les stimuli : il y a facilitation intermodale.

Cependant nous n'avions pas noté l'importance de la congruence sémantique qui peut aussi avoir des conséquences négatives. En effet, lorsque les deux stimuli représentent des choses qui ne sont pas présentées ensemble habituellement, l'effet de facilitation disparaît et un effet d'interférence peut même avoir lieu en cas de conflit entre la composante visuelle et la composante auditive, ce qui conduit à des temps de réponse plus longs. Cela se produit quand le lien sémantique est trop fort entre les deux composantes, comme c'est le cas avec des mots écrits et parlés. Il peut être alors difficile d'ignorer une modalité non pertinente. C'est ce qui se passe dans l'***effet Stroop*** intermodal [Laurienti *et al.* 2004][14]. Dans leur expérience, Laurienti et ses collègues présentent des mots représentant une couleur (exemples : "vert", "bleu", "rouge") visuellement (le mot est écrit), auditivement (le mot est pronocé) ou par une combinaison audiovisuelle. Dans certains cas la même couleur sera représentée dans les deux modalités, dans d'autres on présentera une couleur différente dans chaque modalité. A chaque essai un des deux mots "rouge" ou "bleu" apparaît, dans l'une, l'autre ou les deux modalités. Il s'agit de la cible. Dans les cas où la cible n'est présente que dans une modalité, l'autre modalité présente le mot "vert" qui sert ici de distracteur. Les participants doivent alors indiquer à quelle option de cible ils ont à faire en appuyant sur une touche pour le mot "rouge" et sur une autre touche pour le mot "bleu". Les auteurs ont remarqué un gain de temps lorsque les deux modalités présentent la même couleur et une perte de temps lorsque l'une des deux modalités présente une couleur différente.

Ces effets, positifs et négatifs, dûs à la congruence sémantique, sont présents dans les tâches d'identification d'évènements ou d'objets plus réalistes exposés section 3.4.1.2 b) , que ce soit lorsque les stimuli sont présentés sous forme d'une combinaison de sons d'environnement et d'une représentation visuelle 3D [Suied *et al.* 2007], d'une photographie représentant la source [Yuval-Greenberg et Deouell 2009] ou d'une représentation du contexte environnemental [Ozcan et van Egmond 2009].

14. Initialement l'effet Stroop avait été montré en tant qu'effet intramodal avec la modalité visuelle seulement. La couleur de la police changeait. Il est alors très dur d'associer le mot vert à la couleur qu'il représente lorsqu'il est écrit dans une autre couleur comme dans : vert.

### b) Hypothèse d'unité et crédibilité

Lorsque l'hypothèse selon laquelle les stimuli audio et visuels sont issus d'une même source (ou d'un même évènement) est suffisamment convaincante, ces stimuli seront plus à même d'être fusionnés. La tendance qu'ont les observateurs à accepter que ces stimuli aient une origine spatiotemporelle commune est alors renforcée et il sera plus difficile de discerner lequel des deux est en avance ou en retard lors d'un léger décalage. Cet effet s'observe lorsque l'on présente des vidéos de parole, de musique ou d'évènements du quotidien à des participants [Vatakis et Spence 2008] : le décalage temporel est moins fort quand les deux composantes vont bien ensemble (par exemple le visage d'une femme avec la voix d'une femme), mais plus prononcé dans un cas non cohérent (par exemple le visage d'une femme avec la voix d'un homme, même si les deux prononcent la même phrase). Cette ***hypothèse d'unité*** (*unity assumption*) explique aussi pourquoi la fenêtre d'intégration temporelle (section a) ) est plus large pour des stimuli réalistes comme la parole. [15]

La formulation de la consigne lors des expérimentations est alors primordiale : les participants n'auront pas le même ressenti selon que l'on suggère que les stimuli auditifs et visuels sont à combiner ou qu'ils n'ont pas de rapport. Par exemple « l*es images que vous allez voir peuvent ou non représenter les sources des sons que vous allez entendre* » (voir chapitre 5) ne favorisera pas spécialement l'intégration, tandis que « *les six stimuli seront présentés auditivement, visuellement ou audiovisuellement* » (chapitre 6) laissent sous entendre qu'il y a un lien entre les composantes visuelles et sonores.

### c) Disponibilités et répartition des ressources attentionnelles

Le fait de se concentrer (attention volontaire) sur l'une ou l'autre des modalités influence l'intégration multimodale et le traitement de chaque modalité. C'est l'hypothèse d'***attention dirigée*** (ou sélective) relevée dans [Welch et Warren 1980] : la modalité qui reçoit l'attention domine. Pour mettre en avant l'influence de l'attention volontaire, Yuval-Greenberg et ses collègues [Yuval-Greenberg et Deouell 2009] proposent une tâche d'identification d'animaux présentés toujours par une paire audiovisuelle photographie + son d'animal. Les participants avaient pour consigne de se concentrer sur l'une ou l'autre des modalités. Les résultats de cette étude montrent qu'effectivement la congruence entre l'image et le son influence les temps d'identification : si l'image et le son correspondent il y a un gain en temps de réponse (facilitation) tandis que si les deux éléments ne correspondent pas, une perte de temps supplémentaire (interférence) est observée par rapport à une condition où l'on présente une image neutre, c'est-à-dire sans signification particulière (une mosaïque de parties d'images non reliées les unes aux autres).

15. Nous noterons tout de même que le rôle de l'hypothèse d'unité dans l'intégration multimodale est assez controversée [Spence 2007] puisque les protocoles expérimentaux utilisés pour mettre ce facteur en avant utilisent des mesures sur des paramètres qui peuvent en eux-mêmes influencer l'intégration. Ainsi utiliser des vidéos dont les composantes visuelles et sonores ne sont pas cohérentes implique que l'évolution temporelle entre les deux composantes est différente, or on a vu que la synchronisation temporelle était un facteur important pour l'intégration. De plus il est difficile de définir si l'hypothèse d'unité est réellement un facteur cognitif lié à un processus attentionnel descendant (*top-down*, c'est-à-dire induit par la tâche ou par l'utilisateur) ou si c'est un processus plutôt de type ascendant (*bottom-up*) lié aux propriétés des stimuli [Vatakis et Spence 2008] car les facteurs qui favorisent l'hypothèse d'unité sont aussi des facteurs qui favorisent de façon indépendante l'intégation multimodale.

Ces effets de facilitation et d'interférence dépendent alors de la modalité sur laquelle les participants portent leur attention et sont plus importants pour une attention sélective tournée vers la modalité visuelle.

Ce phénomène d'attention dirigée s'oppose à la notion d'***attention partagée*** dans laquelle il s'agit de suivre et/ou traiter des informations simultanées issues de deux modalités distinctes. La répartission des ressources attentionnelles partagées entre audition et vision semble agir comme une compétition entre ces deux modalités [Robinson et Sloutsky 2004], et se concentrer sur une modalité se fait au détriment de l'autre et provoque un accroissement des temps de réaction dans la seconde modalité.

Pourtant, comme nous l'avons déjà remarqué avec l'effet Stroop audiovisuel (section a) ), il n'est pas toujours possible d'inhiber le traitement d'une modalité, même si elle n'est pas pertinente pour la réalisation de la tâche.

Enfin, on notera un coût cognitif quand on passe d'une modalité à l'autre (*modality swith cost*) [Lukas *et al.* 2010]. Par exemple, dans leur étude [Spence et Driver 1997], Spence et Driver présentent des stimuli cibles visuels ou sonores et montrent que les performances de détection de ces stimuli sont meilleures si deux cibles successives sont présentées dans la même modalité que lorsqu'elles sont présentées dans deux modalités distinctes. En conséquence, si lors d'une évaluation, on est amené à comparer les performances des participants dans une condition audio seul à celles obtenues dans une condition visuel seul, il faut mieux séparer les conditions dans des blocs distincts que de mélanger les essais respectifs à chaque condition dans un même bloc. Cette séparation permet de limiter le coût cognitif du passage d'une modalité à l'autre qui peut biaiser les résultats.

#### d) Connaissances antérieures

Les connaissances antérieures (*priors*) influencent également l'intégration multi-sensorielle. Leur rôle sera d'autant plus fort que ces connaissances sont fiables.

Ainsi, l'habitude de voir et entendre associés une image et un son favorisera l'intégration. On parle de familiarité des co-occurences de stimuli audiovisuels. A l'inverse, présenter simultanément deux stimuli qui ne sont pas présentés ensembles habituellement pourra être perturbant. De même, l'intégration audiovisuelle est plus importante pour des stimuli familiers [Robinson et Sloutsky 2010].

## 3.5 Pistes de recherche et démarche expérimentale

L'objet de ce chapitre était de présenter un état des lieux de la recherche actuelle sur les facteurs humains qui interviennent dans une recherche d'un objet parmi plusieurs. La connaissance de ces facteurs permet en effet une meilleure conception des interfaces d'exploration de documents à base de présentation simultanée. La saillance perceptive et les effets d'attraction involontaire (*pop-out*), observés en vision (section 3.1) et en audition (section 3.3.3), sont déjà utilisés dans des interfaces zoomables (ZUI). Toutefois, alors que les études sur les traits caractéristiques visuels sont très diversifiées et ont mené à de nombreuses interfaces visuelles, peu de traits auditifs ont été étudiés et seul le volume sonore semble être exploité dans les interfaces audio. Se posent alors les questions suivantes :

1) Peut-on trouver d'autres traits caractéristiques en audition ?
2) Le cas échéant, ces traits caractéristiques sont-ils exploitables, et comment le sont-ils dans une interface auditive d'exploration de documents audio ?

Suite à ces questions, nous nous sommes proposée de trouver d'autres traits caractéristiques auditifs, et pour cela nous avons fonctionné par analogie. En particulier nous présenterons dans le chapitre 6 une technique de profondeur de champ sémantique étendue à l'audio à travers la définition d'un flou sonore.

Dans ce chapitre, plusieurs interactions audiovisuelles viennent d'être exposées. Ainsi nous avons pu noter que l'audio est parfois le sens prédominant d'une présentation bimodale. En outre, lors d'une présentation audiovisuelle redondante, un effet de redondance de cibles est observable. Cet effet améliore les temps de réponse ainsi que l'identification d'objets. Il est donc très intéressant de le mettre à profit dans une interface multimodale afin de diminuer les temps de recherche d'un document dans une grande collection. Cela justifie nos recherches sur la conception de stratégies de présentations audiovisuelles.

Un autre effet de redondance a été noté : l'effet de cibles redondantes pour lequel la combinaison de deux traits caractéristiques indiquant une même cible augmente l'effet de *pop-out* visuel. Compte tenu de la compatibilité entre la redondance de cibles et l'effet de cibles redondantes, on peut supposer que la combinaison d'un trait caractéristique visuel et d'un trait caractéristique sonore indiquant un même objet dirigera l'attention sur lui de façon automatique. Par conséquence, la recherche d'un document multimédia devrait être plus rapide pour une présentation bimodale qu'unimodale. Nous testerons cette hypothèse dans le chapitre 4 en reprenant des traits caractéristiques connus en visuel et en audio (taille et volume sonore) sur une tâche appliquée au browsing de vidéos, et dans le chapitre 6 sur un trait caractéristique visuel connu et de son analogue auditif que nous aurons nous-même défini en fonctionnant par analogie (flous visuel et auditif).

Par ailleurs, lorsque plusieurs sources sont entendues simultanément et que l'utilisateur doit se concentrer sur l'une d'elles seulement, le système auditif présente des limites. Comme nous l'avions par ailleurs observé dans le chapitre précédent, peu d'interfaces sonores utilisent une présentation concurrente des données. Nous avons donc eu pour démarche de vérifier que l'ajout d'audio dans un affichage de plusieurs documents audiovisuels était réellement bénéfique, et si oui sous quelle forme et à quelles conditions. Pour cela nous avons systématiquement comparé les rendus audiovisuels obtenus à des rendus visuels afin de mesurer le gain éventuel de l'audio.

## Deuxième partie

# Stratégies de présentation audiovisuelles

Chapitre 4

# Méthodes de présentation par distorsion de l'espace de représentation pour l'exploration de collections multimédia

*The location of visual elements in the UI has a huge impact on how the user interprets information.*
Rick Oppedisan, 2002

## Sommaire



Plusieurs stratégies de présentation, utilisées dans les interfaces d'accès à l'information dans de grandes bases de données, ont été exposées au chapitre 2. Les interfaces ainsi développées sont dites zoomables (ou multi-échelles) et permettent de mettre en avant l'information pertinente en jouant sur des changements d'échelle et la distorsion de l'espace de représentation. Cependant, comme nous l'avons vu dans ce chapitre d'état de l'art, ces stratégies sont soit exclusivement visuelles, soit exclusivement sonores, mais ne sont pas adaptées à la présentation de documents

multimédia. Le chapitre qui suit s'intéresse à l'extension de certaines de ces stratégies de présentation des domaines audio et visuel au domaine de l'audiovisuel. Deux questions se posent alors :

- comment garder une cohérence entre les deux modalités ?
- dans quelle mesure ces outils multimodaux sont-ils applicables à de l'exploration de collection vidéo ?

Pour répondre à ces questions, nous avons dans un premier temps proposé un modèle pour combiner les techniques audio et visuelles de type Pan&Zoom et Focus+Contexte afin d'obtenir des techniques bimodales. Ensuite, à partir de ces techniques, nous avons développé une application de navigation dans une collection de documents audiovisuels. Une étude d'utilisabilité a été menée pour évaluer l'apport d'une présentation bimodale par rapport à une présentation unimodale dans cette application. Les résultats de cette étude ont fait l'objet d'une publication "*Audiovisual Renderings for Multimedia Navigation*" lors de la conférence internationale sur les interfaces sonores (*International Conference on Auditory Display* ICAD'10).

## 4.1 Modélisation et implémentation de stratégies multimodales

Cette section définit un modèle de rendu audiovisuel théorique qui permet de combiner des techniques déjà existantes, mais indépendantes, dans chaque modalité afin d'obtenir des techniques multimodales. Cette modélisation repose sur une mise en relation entre les paramètres du rendu visuel et ceux du rendu sonore. À partir de ce modèle théorique, nous avons implémenté deux stratégies audiovisuelles multi-échelles.

### 4.1.1 Modèles théoriques de stratégies de présentation audiovisuelles

Plusieurs stratégies de présentation visuelle ont été introduites au chapitre 2. Notre but est d'étendre ces stratégies à l'audiovisuel. Or, pour obtenir un rendu audiovisuel intuitif pour l'utilisateur, il faut garder une cohérence entre le rendu graphique et le rendu sonore. Nous avons donc associé certains paramètres visuels à des paramètres sonores par un lien direct (*direct mapping*) :

- la **position des sources sonores** associées aux vidéos est déduite de la **position des objets visuels** à l'écran ;
- le **volume sonore** des vidéos (objets audiovisuels) est relié à leur **taille visuelle**.

Reprenant la taxonomie du chapitre 2, nous nous sommes intéressée à trois méthodes de rendu particulier, à savoir la technique Pan&Zoom (**PZ**), la technique de lentille grossissante en œil-de-poisson ou *Fisheye Lens* (**FL**) et une technique combinant l'affichage bifocal et la transparence : Bifocal+Transparence (**B+T**).

#### 4.1.1.1 Modèle général

Nous avons schématisé en figure 4.1 les relations entre représentation visuelle et représentation sonore pour les trois stratégies mentionnées.

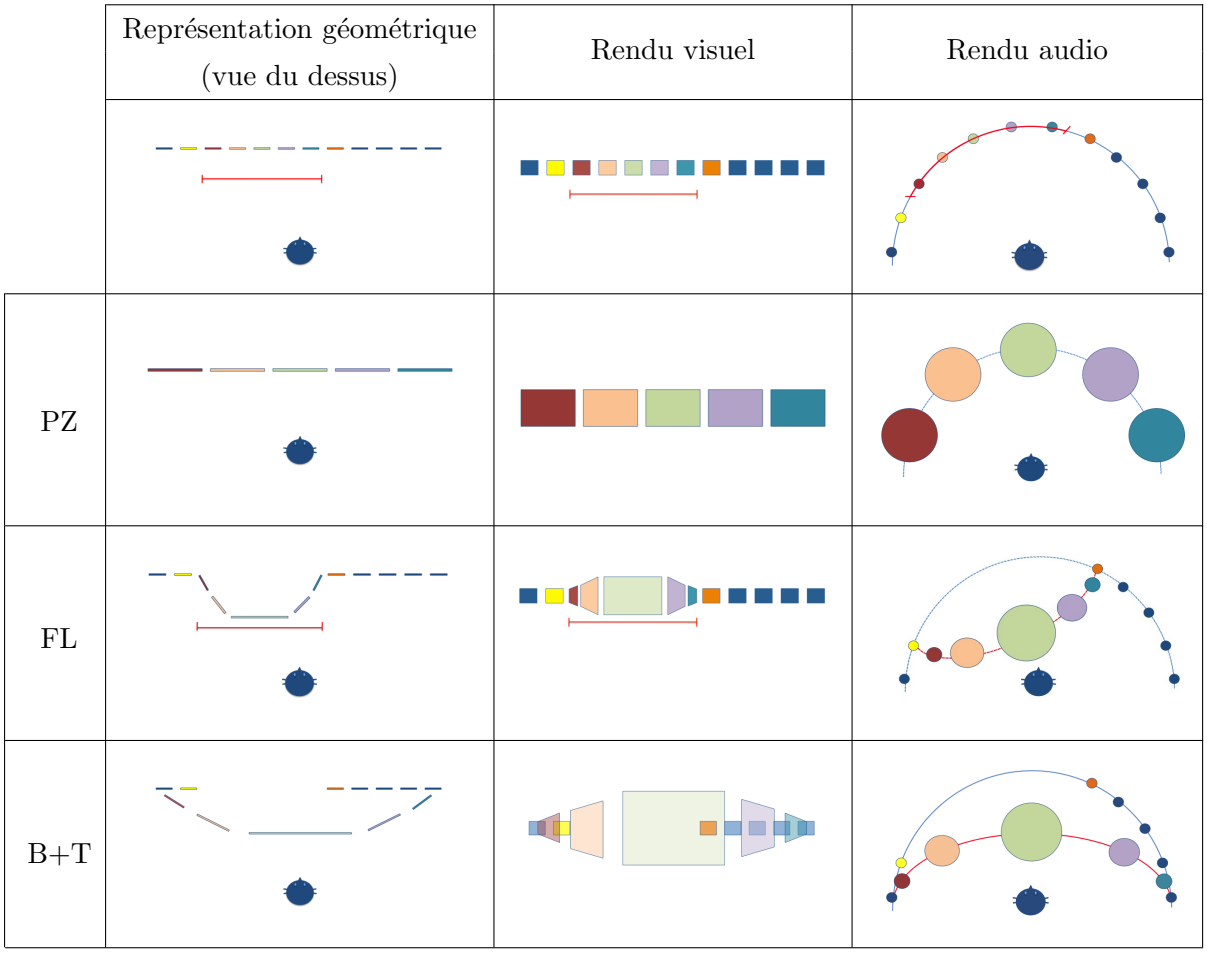



**Figure 4.1.** Schématisation des 3 techniques de rendus : Pan&Zoom (PZ), Lentille Fisheye (FL), et Bifocal + Transparence (B+T). Dans la colonne Rendu audio, les sources sont représentées par des disques. La taille d'un disque indique le volume de la source.

L'extension audiovisuelle des stratégies de présentation que nous proposons ici repose sur la notion d'***espace de représentation***. En effet, la taille et le volume sont des paramètres liés à la distance, et la position fait clairement référence à une notion d'espace. Tous les paramètres (taille + position visuelle, volume + position sonore) que nous avons utilisés pour relier le rendu sonore au rendu graphique sont en fait dépendants de la position apparente des vidéos.

Pour obtenir le rendu audiovisuel à partir des stratégies visuelles, nous avons procédé par ajout du rendu sonore au rendu graphique. Ainsi le rendu visuel est celui que l'on aurait obtenu si l'on avait eu à rendre disponible uniquement la composante visuelle des vidéos : les vidéos sont affichées sur l'écran devant l'utilisateur. Ensuite pour ajouter l'audio, nous avons cherché à tirer profit au maximum de la spatialisation des sources sonores qui facilite la ségrégation de sources sonores simultanées. Alors que la composante visuelle des vidéos sont réparties frontalement et occupent un angle visuel restreint, nous avons espacé les sources sonores sur un arc de cercle le plus large possible, autour de l'auditeur. On peut donc considérer que le rendu graphique, tel qu'on peut l'obtenir sur l'écran, est le développement sur une droite de

l'arc de cercle utilisé pour le rendu audio spatialisé. Le passage de l'espace de représentation graphique à celui de l'audio correspond à une transformation des coordonnées cartésiennes (linéaires) en coordonnées polaires. À partir du rendu graphique (figure 4.1, colonne 2), on peut obtenir une représentation géométrique cartésienne de l'espace global, avec l'utilisateur et les objets à afficher, en prenant le rendu visuel vu du dessus (figure 4.1, colonne 1). Le rendu audio (figure 4.1, colonne 3) est obtenu en enroulant le segment de droite portant les sources graphiques sur un arc de cercle centré sur l'utilisateur. Au final, les positions des composantes audio et visuelle de chaque objet sont donc cohérentes bien que non congruentes. Le lien entre la taille des objets affichés et leur volume sonore est directement inspiré de la perception de la distance au quotidien. Un objet plus grand est perçu comme plus proche, il doit donc sonner plus fort.

#### 4.1.1.2 Spécificité des trois méthodes modélisées

Pour la première méthode reprenant le concept de Pan&Zoom (PZ), il n'y a pas de distorsion : les objets sont affichés avec une taille homogène et le volume sonore est le même pour toutes les sources sonores présentes. En revanche, seuls quelques objets sont affichés. Le reste des objets, c'est-à-dire le contexte, disparaît lorsque le niveau de zoom augmente.

Le design de la lentille en œil-de-poisson (ou *Fisheye Lens* abrégée FL) utilise une distorsion sur la position visuelle. L'équivalent sonore correspond donc à une transformation de l'angle pour le rendu spatialisé des sons. De plus, à l'agrandissement progressif des objets dans le rendu graphique correspond un accroissement progressif du volume dans le rendu sonore. Cette méthode présente l'avantage de fournir la totalité du contexte dans les deux modalités. En revanche, la distorsion dans le rendu graphique peut perturber les utilisateurs. De plus, la distorsion dans le rendu sonore peut être dure à percevoir puisqu'elle dépend d'une variation azimutale assez faible (les sources sont peu déplacées au niveau angulaire). Finalement, les sources sonores sont rapprochées les unes des autres à l'intérieur de la lentille. Il est alors plus difficile de séparer le flux sonore de chaque objet contenu dans la lentille, ce qu'il faut éviter en audio.

Le but de la troisième méthode Bifocal+Transparence (B+T) est, au contraire, d'augmenter la ségrégation des sources sonores dans la zone de focus pour faciliter la tâche de recherche et d'attention sélective de l'utilisateur [Bregman 1990]. L'idée est donc d'étaler les sources sonores du focus le plus possible autour de l'utilisateur. Pour écarter les sources sonores au maximum nous avons utilisé une technique mixte alliant la transparence et la distorsion d'un rendu bifocal de type lentille Fisheye. Ainsi le focus est similaire à celui de la lentille FL mais écarté au maximum et superposé par transparence au contexte. Les sources sonores à l'intérieur de la lentille sont donc plus facilement discernables. Cependant, plusieurs sources (du focus et du contexte) peuvent être entendues à la même position à cause de la superposition.

Pour finir, il est possible de combiner le rendu graphique d'une méthode avec le rendu sonore d'une autre méthode de façon non congruente. Par exemple un rendu visuel sans distorsion PZ pourra être associé à un rendu sonore B+T tel que les sources sonores soient les plus espacées possible. Les rendus audio et graphiques ne sont alors plus congruents mais le lien entre les deux

reste cohérent.

### 4.1.2 Implémentation du rendu visuel

Nous avons implémenté deux modes de rendu graphique : un mode lentille Fisheye (FL) et un mode Pan&Zoom (PZ). Cette section présente les approches informatiques employées pour la mise en place de ces modes de rendu graphique en temps réel.

L'implémentation visuelle doit répondre aux attentes d'une lecture simultanée de plusieurs vidéos avec une distorsion interactive en temps réel. Pour ce faire, nous avons à notre disposition des programmes de type ***shader***. Il s'agit de programmes effectués non pas par le processeur de l'ordinateur (CPU, pour *Central Processing Unit*) mais directement dans la carte graphique (GPU, pour *Graphics Processing Unit*), permettant ainsi de dédier plus de ressources au calcul graphique et de réduire la bande passante sur le bus graphique.

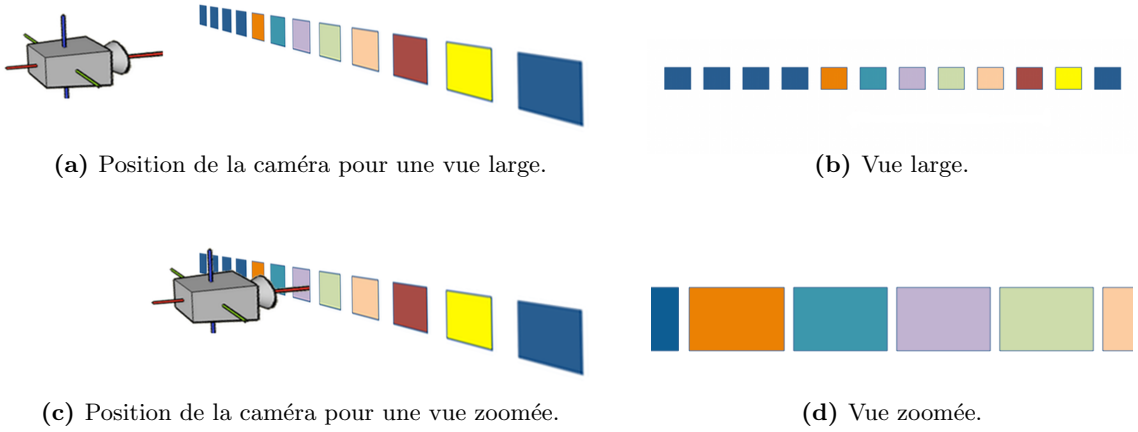
Les *shaders* sont principalement de trois types :
- le *vertex shader* gère l'éclairage et les transformations, en particulier la projection des coordonnées des polygones (primitives permettant de dessiner des objets en 3D) ;
- le *geometry shader* permet de cloner des polygones ou de changer leur disposition ;
- le *fragment shader* (ou *pixel shader*) permet de calculer la couleur de chaque pixel de l'image à afficher. C'est sur ce dernier type de *shader* que nous avons travaillé en imposant au *shader* de choisir la couleur du pixel en fonction du niveau de zoom et de la position du focus.

En pratique, plusieurs langages de programmation sont disponibles pour effectuer du calcul graphique en GPU, notamment le langage Cg (raccourci de *C for Graphics*) et le langage GLSL (pour *OpenGL Shading Language*). Comme nous voulions utiliser Virtual Choreographer (VirChor) [Jacquemin 2004] pour la description des scènes audiographiques et l'exécution temps réel du rendu graphique, programme que nous maitrisions bien, nous avons utilisé le langage Cg qui est le langage de développement de *shaders* utilisé par VirChor.

#### 4.1.2.1 Rendu de type Pan&Zoom

Le mode de rendu du type PZ permet de n'afficher qu'une portion de l'environnement. Cette méthode correspond à la manipulation d'une caméra comme décrit dans les diagrammes *space-scale* de Furnas et Bederson [Furnas et Bederson 1995] (chapitre 2, section 2.1.2). La caméra peut être déplacée selon un axe gauche-droite (axe de la caméra représenté en vert sur la figure 4.2) pour effectuer un mouvement de *pan*, ou selon un axe avant-arrière (axe rouge sur la figure 4.2) pour changer le niveau de détail lors d'un *zoom*. Une part de la vue globale est ainsi agrandie pour occuper toute la fenêtre graphique[1]. Si l'on considère que la vue globale correspond à une vue à l'échelle 1, la vue zoomée est à l'échelle **ZR** (pour *Zoom Ratio*) où **ZR** correspond à l'agrandissement, c'est-à-dire au rapport entre la taille des objets zoomés et la taille des objets dans la vue globale.

1. La notion de fenêtre graphique fait ici référence à la fenêtre d'affichage. Celle-ci peut être de la taille de l'écran (en mode plein écran) mais ne l'est pas obligatoirement.



**(a)** Position de la caméra pour une vue large.

**(b)** Vue large.

**(c)** Position de la caméra pour une vue zoomée.

**(d)** Vue zoomée.

**Figure 4.2.** Position de la caméra par rapport aux objets « vidéos » dans l'environnement. La distance entre la caméra et les objets (colonne de gauche) détermine le rendu visuel (colonne de droite).

#### 4.1.2.2 Rendu de type lentille grossissante

La notion de lentille grossissante provient des distorsions graphiques de type œil-de-poisson explicitées dans [Sarkar et Brown 1994]. Dans leur rendu, le grossissement est maximal au centre de la lentille puis diminue progressivement du centre jusqu'au bord de la lentille pour revenir à une taille normale (échelle=1) à l'extérieur de la lentille. Comme l'échelle varie entre la zone de focus et la zone de contexte, cette méthode est parfois qualifiée de méthode à échelle variable [Harrie *et al.* 2002]. De plus, cette méthode est aussi dite à base de distorsion [Leung et Apperley 1994] car les objets situés dans la zone de changement d'échelle apparaissent déformés (par exemple, un rectangle pourra apparaître sous forme de trapèze). L'avantage de cette technique est que l'on conserve l'affichage du contexte, contrairement à la méthode précédente Pan&Zoom, et que le focus ne masque pas de partie du contexte, contrairement à une approche bifocale sans transparence.

En pratique, nous avons décidé de laisser dans la lentille une zone centrale (disque de rayon $\mathbf{rad_{int}}$) dans laquelle le focus est homogène et présenté sans distorsion, comme dans la technique présentée par [Yamamoto *et al.* 2009]. La fenêtre graphique est alors partitionnée en trois zones en fonction de la distance $\mathbf{r}$ entre un point de l'image et le centre de la lentille (voir figure 4.3) :

- pour $\mathbf{r} > \mathbf{rad_{ext}}$, on est à l'extérieur de la lentille : la vue est large, homogène, avec un niveau de détail faible, l'échelle est minimale (échelle=1) ;
- pour $\mathbf{r} < \mathbf{rad_{int}}$, on est à l'intérieur de la lentille : la vue est resserrée, homogène, avec un niveau de détail fin, l'échelle est maximale (échelle=**ZR**) ;
- entre les deux $\mathbf{rad_{int}} < \mathbf{r} < \mathbf{rad_{ext}}$, on est dans une zone de transition qui permet de compenser l'effet de grossissement de l'intérieur de la lentille : les objets sont distordus.

Nous avons implémenté successivement deux *shaders* pour obtenir un rendu graphique de cette lentille grossissante qui nous semble convenable.

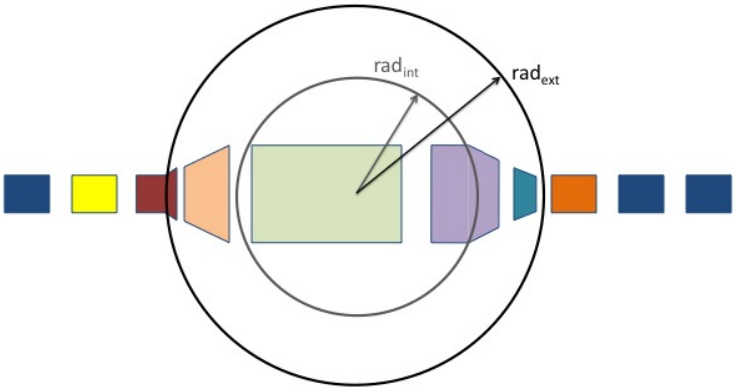



**Figure 4.3.** Schéma du rendu visuel obtenu en fonction des trois zones définies par les rayons interne $\mathbf{rad_{int}}$ et externe $\mathbf{rad_{ext}}$ de la lentille. En pratique une distorsion courbe apparaît sur les bords des trapèzes dans la zone de transition. Cette distorsion n'est pas indiquée ici par soucis de schématisation.

#### a) Première proposition

Notre première proposition de *shader*, que nous présentons dans ce paragraphe, avait été développée en collaboration avec Matthieu Courgeon du LIMSI-CNRS. Elle s'appuie sur une procédure en deux passes, c'est-à-dire que le rendu graphique est calculé deux fois. Il s'agit en effet de la façon la plus simple, et la moins coûteuse en temps de calcul, d'obtenir un rendu de type lentille grossissante avec un *shader*. La première fois que le rendu est calculé, la scène graphique est placée horizontalement, parallèle au plan de projection, et projetée en perspective. Le rendu graphique obtenu, une vue d'ensemble à l'échelle 1, est enregistré dans une texture *TextNorm* (l'équivalent d'une capture écran sans que l'image ne soit affichée). Cette texture est ensuite déformée grâce à un *fragment shader* en suivant la stratégie œil-de-poisson : les pixels situés à l'intérieur de la lentille sont étirés afin d'apparaître plus larges et les pixels au bord de la lentille sont au contraire comprimés. En réalité, on ne change pas la taille des pixels, mais on transforme les coordonnées de texture des pixels de l'image capturée *TextNorm* afin d'obtenir la texture à afficher sur toute la surface de la fenêtre graphique. Pour donner l'impression d'un pixel plus grand, on peut par exemple donner la même couleur à deux pixels voisins en divisant par deux les coordonnées de texture sur cette zone. Comme la lentille que nous avons définie est circulaire, nous avons caractérisé les pixels par leur distance par rapport au centre de la lentille $\mathbf{r1}$ et $\mathbf{r}$, respectivement la distance dans la texture à l'échelle 1 *TextNorm* et la distance dans la fenêtre de rendu graphique ou texture à échelle variable.

La déformation que nous avons choisie d'appliquer à la texture *TextNorm* pour obtenir le rendu graphique est définie par une **fonction de transformation de coordonnées** $\mathbf{r} = \mathbf{T}(\mathbf{r1})$ telle que présentée en équation 4.1. Elle dépend des paramètres de rayons interne et externe de la lentille, $\mathbf{rad_{int}}$ et $\mathbf{rad_{ext}}$, et de l'échelle maximale $\mathbf{ZR}$ (ou niveau de zoom) que l'on veut atteindre au centre de la lentille. Puisqu'avec cette méthode de lentille grossissante il n'y a pas de masquage, tous les pixels de la texture initiale seront affichés. La fonction de transformation

de coordonnées permet de définir sur quel pixel de position **r** du rendu graphique sera affiché le pixel de position **r1** de la texture initiale à l'échelle 1. En pratique, comme le *shader* calcule, pour chaque pixel de coordonnée **r**, ce qu'il doit afficher de la texture *TextNorm*, c'est la fonction réciproque $\mathbf{r1} = \mathbf{T^{-1}}(\mathbf{r})$ que l'on code (équation 4.2).

$$r = T(r1) = \begin{cases} ZR \times r1 & \text{si } r1 \leq \dfrac{rad_{int}}{ZR}, \\ \dfrac{(ZR \times rad_{ext} - rad_{int}) \times r1}{(rad_{ext} - rad_{int}) - (1 - ZR) \times r1} & \text{si } \dfrac{rad_{int}}{ZR} \leq r1 \leq rad_{ext}, \\ r1 & \text{si } r1 \geq rad_{ext}. \end{cases} \tag{4.1}$$

$$r1 = T^{-1}(r) = \begin{cases} \dfrac{r}{ZR} & \text{si } r \leq rad_{int}, \\ \dfrac{(rad_{ext} - rad_{int}) \times r}{(1 - ZR) \times r + ZR \times rad_{ext} - rad_{int}} & \text{si } rad_{int} \leq r \leq rad_{ext}, \\ r & \text{si } r \geq rad_{ext}. \end{cases} \tag{4.2}$$

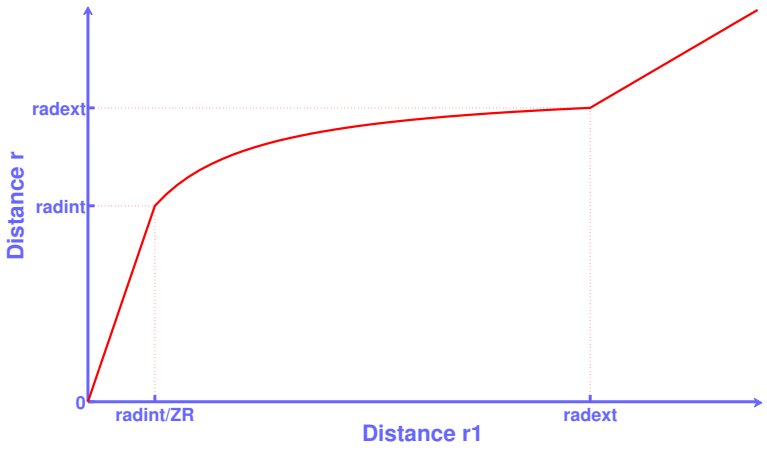



**Figure 4.4.** Courbe de la transformation de coordonnées qui permet de passer de la texture initiale à l'échelle 1 au rendu graphique à échelle variable.

Lorsque l'on applique la fonction de transformation à la texture *TextNorm*, tout se passe comme si on modifiait l'échelle d'affichage. On peut alors définir une **fonction d'échelle** pour décrire le rapport entre la largeur $dr$[2] obtenue dans l'espace image (rendu graphique) et la largeur initiale *dr1* dans l'espace texture en fonction de la distance **r1** au centre de la lentille (voir équation 4.3 et figure 4.5). Cette fonction d'échelle est donc la dérivée de la fonction de transformation. Elle n'est pas constante sur l'ensemble du rendu graphique d'où le nom de méthode à échelle variable.

Lors de la deuxième passe, c'est la nouvelle texture, obtenue par déformation, qui est affichée sur l'ensemble de la fenêtre graphique.

Tenant compte des travaux de Zanella et ses collaborateurs [Zanella *et al.* 2000], un reflet

2. *dr* représente ici un petit déplacement de la distance **r**.

a été ajouté pour renforcer l'effet de relief de la lentille, et un tour blanc permet de délimiter clairement le début de la distorsion et de situer la lentille pour un faible niveau de distorsion.

$$Echelle(r1) = \frac{dr}{dr1}(r1) = T'(r1) = \begin{cases} ZR & \text{si } r1 \leq \dfrac{rad_{int}}{ZR}, \\ \dfrac{k}{A \times r^2 + B \times r + C} & \text{si } \dfrac{rad_{int}}{ZR} < r1 < rad_{ext}, \\ \quad avec\ k = (ZR \times rad_{ext} - rad_{int})(rad_{ext} - rad_{int}), & \\ \quad A = (1 - ZR)^2, & \\ \quad B = -2(1 - ZR)(rad_{ext} - rad_{int}), & \\ \quad et\ C = (rad_{ext} - rad_{int})^2 & \\ 1 & \text{si } r \geq rad_{ext}. \end{cases} \tag{4.3}$$


**Figure 4.5.** Courbe de la fonction d'échelle appliquée à la texture à déformer *TextNorm* décrite dans l'équation 4.3.

Cette méthode présente un défaut majeur d'un point de vue visuel du fait que la déformation a lieu directement sur les pixels : à l'intérieur de la lentille et pour un facteur d'échelle supérieur à 1, un effet non désiré de pixellisation apparaît.

#### b) Lentille Fisheye sans pixellisation

Pour pallier le problème de pixellisation de la proposition précédente, nous avons décidé de fonctionner en trois passes (le rendu graphique est calculé trois fois avant d'être affiché) au lieu de deux. Le temps de calcul est plus long. La première passe enregistre comme précédemment une vue d'ensemble à l'échelle 1 dans une texture nommée *TextNorm*. La seconde passe enregistre, dans une autre texture *TextZoom*, une vue en projection en perspectives à l'échelle supérieure **ZR**, en rapprochant la caméra des vidéos (figures 4.2d et 4.2c). Cette texture correspond à la texture qui serait affichée en mode Pan&Zoom (PZ) à cette même échelle.

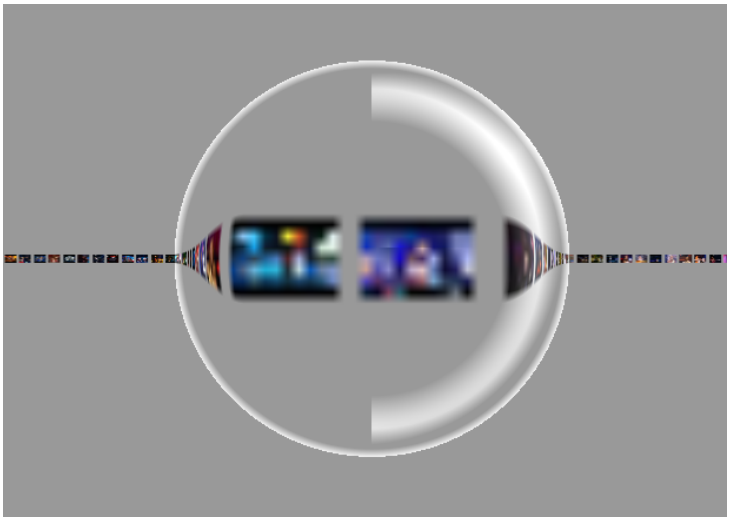


**Figure 4.6.** Effet de pixellisation observé au centre de la lentille. Le niveau de zoom est volontairement accentué pour que l'effet de pixellisation soit clairement visible sur cette image.

La troisième passe consiste à sélectionner et transformer les parties de chaque texture qui vont être utilisées pour construire la texture finale visible à l'écran. Cette sélection dépend du niveau de zoom **ZR** et de la fonction de changement de coordonnées que l'on doit appliquer à chaque texture. Dans la zone de focus homogène (rayon $\mathbf{rad_{int}}$), on utilise la vue zoomée *TextZoom*. Il n'y a pas de changements de coordonnées de texture, la taille des pixels n'est pas modifiée, donc il n'y a pas de pixellisation. A l'extérieur de la lentille (de rayon $\mathbf{rad_{ext}}$), la texture normale *TextNorm* est utilisée, là encore à l'échelle de cette texture. Dans la zone de transition (entre $\mathbf{rad_{int}}$ et $\mathbf{rad_{ext}}$), il faudrait utiliser la texture d'échelle supérieure *TextZoom* pour éviter la pixellisation. Malheureusement pour les forts niveaux de zoom (typiquement $\mathbf{ZR} = 20$), cette texture ne couvre pas l'ensemble de la zone intermédiaire. Nous avons utilisé cette texture sur la zone comprise entre $\mathbf{rad_{int}}$ et $(\mathbf{rad_{int}} + \mathbf{rad_{ext}})/2$ pour ne pas avoir de pixellisation pour les vidéos dont l'échelle est proche de l'échelle maximale. La fonction qui décrit quel pixel positionné à distance **r1** du centre de la lentille dans la texture *TextZoom* doit être appliqué au pixel positionné à distance **r** du centre de la lentille dans la fenêtre graphique, fonction réciproque de la fonction de changement de coordonnées, est définie par l'équation 4.4. Pour la partie de la zone intermédiaire la plus éloignée du centre, entre $(\mathbf{rad_{int}} + \mathbf{rad_{ext}})/2$ et $\mathbf{rad_{ext}}$, c'est la texture *TextNorm* qui est utilisée et déformée selon la fonction en équation 4.5, alors équivalente à la fonction réciproque de la fonction de changement de coordonnées utilisées dans la première proposition de *shader*, à deux passes, vue précédemment (équation 4.2, p. 68). Comme dans cette première proposition, nous avons ajouté un bord blanc pour marquer la frontière de la lentille et ainsi permettre à l'utilisateur de facilement la localiser même dans le cas où le niveau de zoom est très faible.

Passage des coordonnées **r** du rendu graphique aux coordonnées **r1** de la texture *TextZoom*

$$r1 = \begin{cases} r & \text{si } r \leq rad_{int}, \\ \dfrac{r \times ZR \times (rad_{int} - rad_{ext})}{(ZR-1) \times r + radint - ZR \times rad_{ext}} & \text{si } rad_{int} \leq r \leq \dfrac{rad_{int} + rad_{ext}}{2}. \end{cases} \tag{4.4}$$

Passage des coordonnées **r** du rendu graphique aux coordonnées **r1** de la texture *TextNorm*

$$r1 = \begin{cases} \dfrac{(rad_{ext} - rad_{int}) \times r}{(1 - ZR) \times r + ZR \times rad_{ext} - rad_{int}} & \text{si } \dfrac{rad_{int} + rad_{ext}}{2} \leq r \leq rad_{ext}, \\ r & \text{si } rad_{ext} \leq r, \end{cases} \tag{4.5}$$

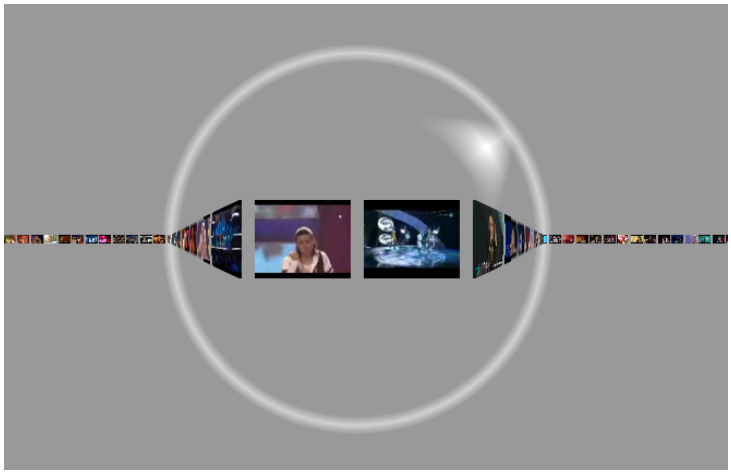


**Figure 4.7.** Rendu visuel obtenu avec la lentille finale en trois passes.

### 4.1.3 Implémentation du rendu sonore

De façon à accentuer la séparation spatiale entre les sources et ainsi faciliter la ségrégation des sources sonores, nous avons implémenté un rendu audio de type Bifocal+Transparence (B+T) dont les paramètres sont le volume sonore et la position azimutale des sources. Nous avons adapté ce mode de rendu pour qu'il puisse être combiné indifféremment avec le rendu visuel PZ ou le rendu visuel FL. Dans le rendu visuel PZ il n'y a pas de distorsion (changement d'échelle non homogène) visuelle, nous avons donc cherché à minimiser les distorsions appliquées à l'audio. Il n'y a donc aucun changement local d'échelle pour l'azimut (**az**) dans ce mode (équation (4.7)). Cependant, bien que nous ayons dû appliquer, en théorie, une augmentation homogène du volume sur les sources du focus, nous avons conservé une distorsion, faible, sur le volume (**vol**) afin de réduire le nombre de sources saillantes entendues simultanément (équation (4.6)). La distorsion de volume est similaire à celle employée pour la méthode FL (équation (4.8)), mais avec un rayon de lentille $\mathbf{rad_{ext}}$, égal à la largeur de la fenêtre de rendu graphique.

Pour être associé au rendu visuel FL, le rendu sonore repose cette fois sur un changement d'échelle non homogène à la fois de la position azimutale et du volume sonore. La figure 4.9 présente la courbe de distorsion du volume sonore de chaque objet en fonction de sa distance au centre de la lentille dans l'espace image. Les valeurs du volume et de la position azimutale des sources audio sont données par les équations (4.8) pour le volume **vol** et (4.9) pour l'azimut apparent dans l'espace sonore ($\mathbf{az_{bis}}$). Dans ces équations :

– **c** est une constante, obtenue empiriquement, qui caractérise l'atténuation non linéaire entre la zone de focus (centre de la lentille) et le contexte (extérieur de la lentille) ;

- **az** est l'azimut dans l'espace sonore à l'échelle 1 (avant zoom) ;
- **r** représente la distance visuelle apparente entre le centre de l'objet et le centre de la lentille dans la fenêtre graphique ;
- **ZR** est l'échelle maximale, au centre de la lentille, qui définit le niveau de zoom ;
- $\mathbf{v_{min}}$ est le volume sonore minimal, c'est-à-dire celui qu'ont les sources quand il n'y a pas de grossissement (**ZR** égal à 1) ou quand la source est située à l'extérieur de la lentille. Si l'on veut éteindre les sources en dehors de la lentille, on peut poser $\mathbf{v_{min}} = 0$ ;
- $\boldsymbol{\alpha_{int}}$ et $\boldsymbol{\alpha_{max}}$ représentent réciproquement l'azimut des sources sur le périmètre interne de la lentille (cercle de rayon $\mathbf{rad_{int}}$ dans l'espace image correspondant à la fenêtre graphique) et sur le périmètre extérieur de la lentille (cercle de rayon $\mathbf{rad_{ext}}$).

Equations pour le rendu audio de la méthode PZ :

$$vol = \begin{cases} v_{min} + log(ZR) & \text{si } r \leq rad_{int}, \\ v_{min} + log(ZR) \times e^{-c \times |r - rad_{int}|} & \text{si } rad_{int} \leq r \leq rad_{ext}, \\ v_{min} & \text{si } r > rad_{ext}. \end{cases} \tag{4.6}$$

$$az_{bis} = az \tag{4.7}$$

Equations pour le rendu audio de la méthode FL : FL distorsion

$$vol = \begin{cases} v_{min} + log(ZR) & \text{si } r \leq rad_{int}, \\ v_{min} + log(ZR) \times e^{-c * |r - rad_{int}|} & \text{si } rad_{int} \leq r \leq rad_{ext}, \\ v_{min} & \text{si } r \geq rad_{ext}. \end{cases} \tag{4.8}$$

$$az_{bis} = \begin{cases} r \times \dfrac{\alpha_{int}}{rad_{int}} & \text{si } |dz| \leq rad_{int}, \\ A \times r + B & \text{si } rad_{int} < r < rad_{ext} \\ avec\ A = \dfrac{\alpha_{max} - \alpha_{int}}{rad_{max} - rad_{int}}\ et\ B = sg(r), & \\ az & \text{si } r \geq rad_{ext}. \end{cases} \tag{4.9}$$

Parmi les différentes techniques de restitution spatialisée du son existantes (voir chapitre 3, page 40), nous avons choisi la technique d'Ambisonic virtuel. Cette technique nous permet ainsi de spatialiser beaucoup de sources sonores simultanées (une centaine dans l'expérience qui suit), sans augmenter le coût de calcul, et rend possible un rendu sur casque pour le grand public. Cependant, le système de diffusion peut être remplacé sans plus de calcul par un système plus immersif comme de l'Ambisonic ou, avec des calculs différents mais le même concept, de la WFS.

### 4.1.4 Architecture logicielle globale

L'architecture globale de l'application repose sur une suite d'outils appelée ***SceneModeler*** qui avait été précédemment été développée au LIMSI pour de la création de scène virtuelle multimédia [Bouchara 2008]. Le *SceneModeler* est constitué de deux parties : un moteur de rendu

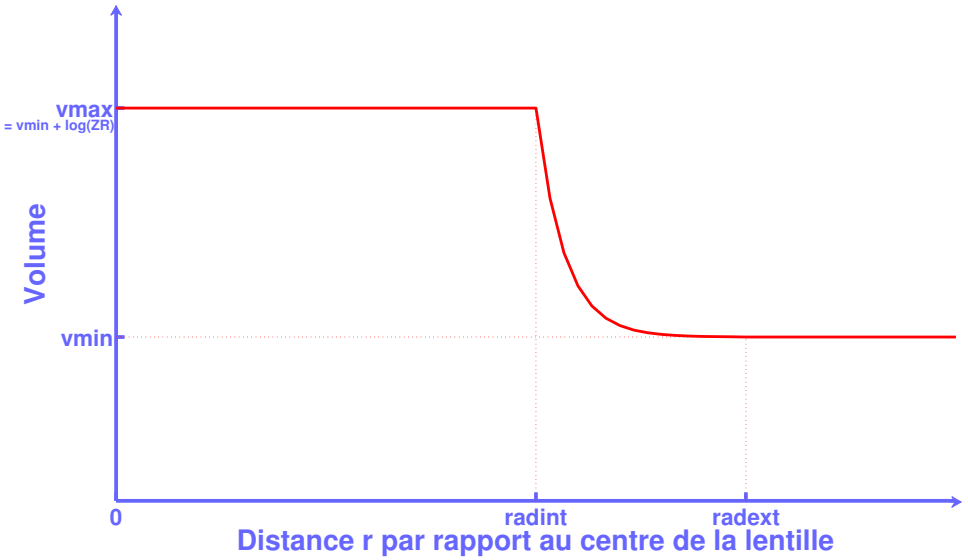



**Figure 4.8.** Courbe de changement d'échelle du volume utilisé dans la méthode audio FL pour calculer le volume des sources sonores en fonction de leur position visuelle dans la fenêtre de rendu graphique.

graphique 3D, appelé ***Virtual Choreographer***, et un spatialisateur temps réel, ici programmé pour le langage Max/MSP. La communication entre le descripteur de scènes et le spatialisateur se fait par protocole UDP avec des messages encodés en OSC [Wright *et al.* 2003].

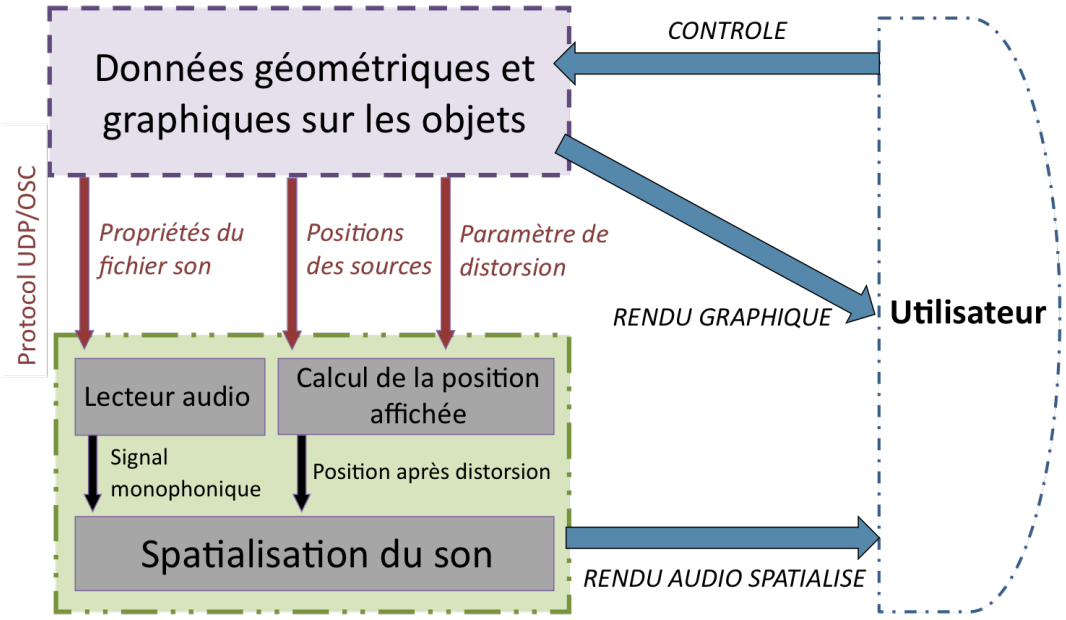



**Figure 4.9.** Schéma de l'architecture générale.

## 4.2 Protocole expérimental pour une évaluation d'utilisabilité sur une recherche de vidéos

Nous avons mis en place une évaluation d'utilisabilité de stratégies de présentation audiovisuelles sur une application réaliste de recherche de vidéos. Le but principal de cette évaluation est de mesurer l'éventuelle contribution de l'audio pour explorer une base de données audiovisuelles. Nous avons, d'une part, comparé des rendus audiovisuels avec des rendus seulement visuels. D'autre part, nous avons testé deux modes de rendus, Fisheye Lens FL et Pan&Zoom PZ, soit audiovisuel soit visuel seul, afin de voir si l'influence de l'audio pouvait dépendre du mode de rendu choisi.

### 4.2.1 Description de la tâche

La tâche expérimentale que devaient réaliser les participants dans cette expérience était inspirée des travaux menés sur la notion de pointage en IHM [Pietriga *et al.* 2007]. Ces travaux visent notamment à comparer différentes techniques de pointage pour des rendus graphiques multi-échelles comme les techniques Pan&Zoom et les lentilles grossissantes. Ainsi l'utilisateur devait, à chaque essai, regarder un clip vidéo particulier puis naviguer au sein d'une collection de 100 vidéos grâce aux techniques proposées afin de retrouver la vidéo cible dans l'ensemble de la collection. Chaque essai était divisé en trois phases représentées schématiquement dans un story-board en figure 4.10 :

1) présentation de la vidéo cible ;

2) exploration de la collection pour trouver la cible ;

3) sélection/validation de la cible.

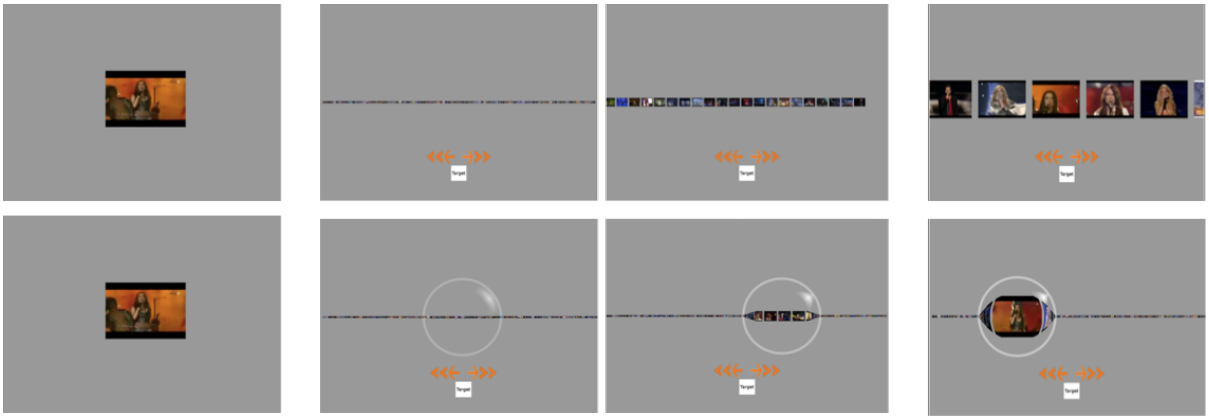


**Figure 4.10.** Les trois phases d'un essai pour chacune des deux méthodes PZ (haut) et FL (bas) : le participant regarde la vidéo cible, explore la collection à l'aide des outils de grossissement, retrouve et sélectionne la cible.

### 4.2.2 Hypothèses

D'après l'état de l'art (chapitres 2 et 3), associer un rendu audio au rendu visuel permet de transmettre des informations redondantes qui renforcent ainsi le retour visuel, et/ou de transmettre des informations supplémentaires pour compléter l'information visuelle. Par conséquent, nous avons supposé que la combinaison d'un rendu audio redondant et complémentaire améliorerait la navigation et l'accès à l'information au sein d'une collection non organisée de documents audiovisuels.

De plus, nous pensions que la technique de rendu choisie pourrait avoir un effet sur le temps de recherche et sur l'utilité de l'audio. En effet, l'ajout d'un élément sonore pour chaque objet visuel présenté peut entraîner une charge cognitive supplémentaire trop importante à traiter. Ainsi nous supposions qu'un rendu sonore plus ciblé (comme celui utilisé avec la lentille FL), c'est-à-dire avec plusieurs sources sonores simultanées mais seulement quelques unes détaillées car plus pertinentes, serait plus utile qu'un rendu moins focalisé (comme le rendu audio PZ) .

### 4.2.3 Paramètres des modes de rendu testés

Le rendu graphique standard, à l'échelle 1, est celui qui permet d'afficher, en même temps à l'écran, l'ensemble des 100 vidéos alignées. Ensuite, nous avons proposé deux modes de rendus, FL et PZ, pour zoomer, c'est-à-dire changer l'échelle.

Le rendu graphique de la méthode FL était tel que présenté section 4.1.2.2, c'est-à-dire basé sur une distorsion de type lentille en œil-de-poisson. Le rayon externe de la lentille était défini par $\mathbf{rad_{ext}} = 178\ px$ et son rayon interne $\mathbf{rad_{int}} = 2/3 * \mathbf{rad_{ext}}$, pour une fenêtre graphique de 1270×940 pixels, toujours en plein écran. Pour le rendu multimodal audiovisuel, le rendu audio correspondait au rendu audio Bifocal+Transparence tel que présenté en figure 4.1 adapté à la méthode FL et dans les équations 4.8 et 4.9. En dehors de la lentille, les sources sonores étaient spatialement réparties entre $-\boldsymbol{\alpha_{max}}$ et $\boldsymbol{\alpha_{max}} = \pm 90°$. Les sources sonores placées à l'intérieur du contour externe ($\mathbf{r} < \mathbf{rad_{ext}}$) de la lentille étaient réparties entre $-\boldsymbol{\alpha_{max}}$ et $\boldsymbol{\alpha_{max}} = \pm 90\ °$, avec les sources appartenant à la zone interne de la lentille ($\mathbf{r} < \mathbf{rad_{int}}$) réparties cette fois entre $-\boldsymbol{\alpha_{int}}$ et $\boldsymbol{\alpha_{int}} = \pm 70\ °$. Les sources internes à la lentille et les sources externes à la lentille sont donc superposées spatialement. La distinction entre ces sources se fait donc surtout au niveau du volume sonore. Avec le niveau de déformation le plus élevé sur cette technique de rendu audio ($\mathbf{ZR} = 20$), seules trois sources simultanées différentes peuvent être entendues distinctement, à des positions spatiales séparées. L'exemple vidéo 4.1 présente un essai (présentation d'une cible + phase de recherche) avec le rendu audiographique obtenu pour la méthode FL. 4.1

Pour le rendu visuel de PZ, aucune distorsion n'a été utilisée. Le rendu audio était lui basé sur la même distorsion que pour la technique FL à l'exception des valeurs de rayons. En effet, nous avons opté pour un compromis entre la non distorsion imposée par le visuel (pour garder au maximum une cohérence audio/visuel) et la nécessité de limiter le volume sonore et le nombre de sources sonores simultanées : le rayon de la lentille audio PZ est alors très large. Pour valeur numérique, nous avons pris un rayon de lentille égale à la largeur de la fenêtre graphique, c'est-à-dire $\mathbf{rad_{ext}} = 1270$ px à l'échelle 1 ($\mathbf{ZR} = 1$). Plus de sources sonores pouvaient être entendues simultanément avec PZ qu'avec FL. Même au niveau de zoom le plus élevé ($\mathbf{ZR} = 20$) jusqu'à

six ou sept sources étaient entendues. Le rendu obtenu pour la méthode PZ est présenté dans
4.2 l'exemple vidéo 4.2.

### 4.2.4 Contrôle

Pour limiter le nombre de facteurs pouvant influencer les performances, nous avons cherché à ce que le positionnement du focus se fasse de la même façon quelle que soit la condition. Ainsi que ce soit pour FL ou pour PZ, les deux méthodes sont contrôlées de la même manière, à l'aide de deux boutons en forme de flèches situés dans la fenêtre graphique. Dès que la souris passe au-dessus de l'un de ces boutons le centre du focus est déplacé à droite ou à gauche. Ce type de contrôle est moins performant que le contrôle habituel des méthodes PZ (barre de défilement ou cliquer/glisser) et moins performant que celui des méthodes FL (rattachement du centre de la lentille au curseur de la souris). Cependant n'étant utilisé ni pour la méthode PZ ni pour la méthode FL, nous avons choisi ce mode car il ne semblait pas favoriser une méthode par rapport à l'autre. En effet, les difficultés de pointage associées aux contrôles habituels des méthodes FL et PZ sont différentes et entraînent des contraintes distinctes que nous voulions limiter. Ces difficultés de pointage ont fait le sujet de plusieurs études, notamment pour la méthode FL [Gutwin 2002].

### 4.2.5 Collection de vidéos

De façon à situer cette étude dans un contexte le plus réaliste possible, nous avons utilisé un corpus de vidéos provenant d'Internet. Ainsi le corpus était constitué de 100 vidéos musicales extraites des finales du Concours Eurovision de la Chanson de 2005 à 2008 (figure 4.11). Chaque vidéo était donc une sorte de clip musical avec la performance d'un chanteur ou d'un groupe de chanteurs différent. Pour chaque vidéo nous avons extrait un clip d'une durée de 10 secondes correspondant à une phrase musicale. Durant les phases de recherche, ces clips étaient lus en boucle. Les différents clips vidéo étaient différentiables par les propriétés visuelles des chanteurs, leurs caractéristiques vocales, ainsi que par les propriétés de la musique et de la lumière fournissant plusieurs indices pour l'identification. Nous avons choisi ce type de vidéos pour l'équilibre entre la quantité d'informations apportées par le son et celles apportées par la vision. De plus, la redondance d'information entre ces deux modalités améliore l'identification. En outre, ce choix a été motivé par l'homogénéité de la qualité des vidéos car elles étaient toutes issues du même programme de télévision.

Pour la composante visuelle, les vidéos étaient enregistrées au même format de 160×120 px et affichées, avant agrandissement, à une taille de 11×8 px. Sur le plan technique, les vidéos étaient juxtaposées en une mosaïque de vidéos afin qu'il n'y ait qu'un seul fichier vidéo à charger (et à appliquer en tant que texture) pour l'ensemble de l'expérience. La figure 4.11 est une image extraite de cette vidéo mosaïque.

Les bandes-son associées aux vidéos étaient séparées puis stockées comme des fichiers audio monophoniques (en ne gardant que le canal gauche), en 44.1 kHz et 16 bits. Ces bandes-son contenaient le chant et l'instrumentation. Pour spatialiser de manière cohérente les sons par rapport aux images, nous avons fait correspondre le centre de la source sonore au centre de l'objet visuel dans l'espace géométrique.

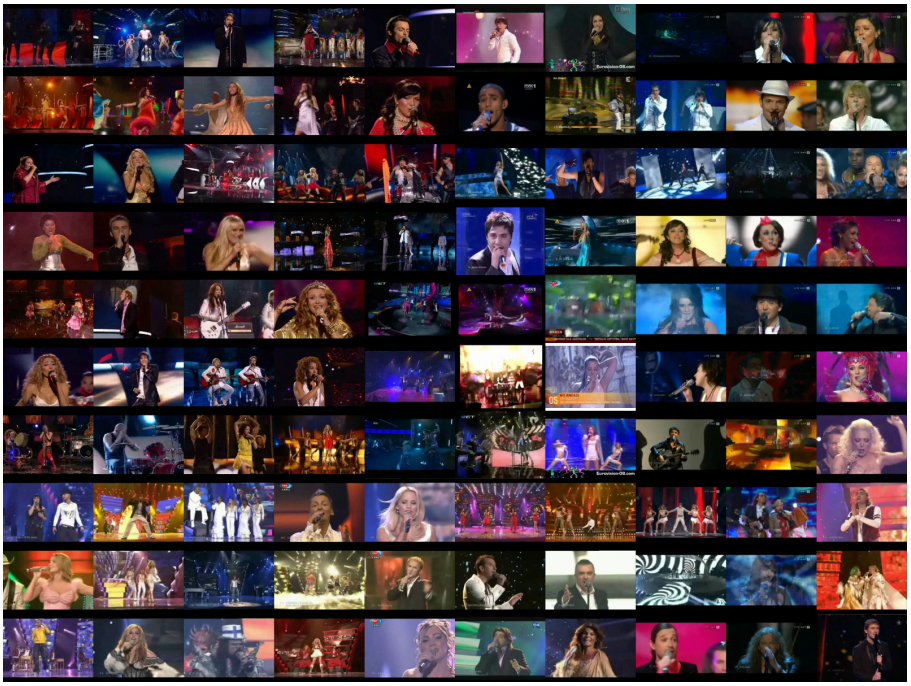
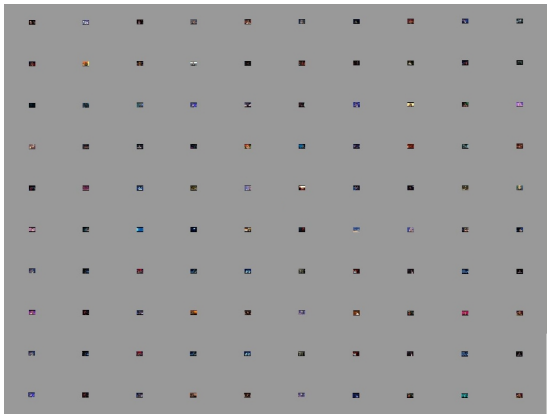


**Figure 4.11.** Mosaïque composée de 100 images issues de chacun des stimuli vidéos.

### 4.2.6 Disposition des vidéos

Nous avions pensé au départ à disposer l'ensemble des vidéos sur une grille, à l'image du Wall de *BlinkX*. Cependant, ce type de présentation en grille avait selon nous deux défauts. D'une part, nous voulions tester les modes de rendus FL et PZ dont le principal intérêt est la fonction d'agrandissement (*zoom*). Il fallait donc obliger les participants à se servir de cette fonctionnalité et pour cela nous devions afficher les vidéos avec une taille suffisamment petite au départ (figure 4.12a). En conséquence, il aurait fallu des milliers de vidéos pour remplir l'espace écran. Malheureusement cela aurait nécessité des sessions expérimentales trop longues car plus il y a de vidéos présentées simultanément plus l'utilisateur met de temps en moyenne à explorer la collection. Inversement, si l'on remplit l'espace écran avec les 100 vidéos à notre disposition, les vidéos sont alors suffisamment grandes pour être distinguées sans avoir recours à un agrandissement, et les techniques multi-échelles Pan&Zoom et Fisheye Lens n'ont pas besoin d'être utilisées (figure 4.12b).

D'autre part, nous avions prévu de spatialiser les sons de façon à mieux les distinguer les uns des autres tout en gardant une cohérence avec le positionnement graphique. Or les techniques de rendus 3D audio actuelles sont relativement limitées en ce qui concerne l'élévation et sont beaucoup plus efficaces en spatialisation horizontale. C'est pour ces deux raisons que nous avons préféré placer les vidéos sur une ligne horizontale.



**(a)** Si les vidéos sont petites, il y a trop d'espace entre elles.

**(b)** Sans écartements, les vidéos ne requièrent pas une fonction de zoom.

**Figure 4.12.** Essais de disposition en grille.

### 4.2.7 Plan d'expérience et procédure

Nous avons utilisé un plan factoriel 2×2 avec pour facteurs : 2 conditions de modalité (V ; AV) et 2 méthodes (PZ ; FL).

Après une session d'apprentissage (sur l'ensemble de ces 4 conditions), l'expérience réelle était divisée en quatre sessions correspondant aux quatre conditions du plan factoriel : AV-FL ; V-FL ; AV-PZ ; V-PZ. L'ordre des sessions était contrebalancé entre les participants par un carré latin. Chaque session se composait de 15 essais, avec à chaque essai un nouveau clip vidéo choisi au hasard comme cible. L'ordre des cibles au sein d'une session était aussi aléatoire. Après chaque session, les participants étaient invités à commenter la condition testée et à évaluer pour chaque condition : l'efficacité, l'adaptabilité, et la difficulté de la combinaison modalité/méthode employée. Après l'expérience, les participants remplissaient un nouveau questionnaire pour donner leur méthode et leur modalité préférées. L'expérience totale durait environ une heure et demie par participant. Les participants étaient encouragés à prendre des pauses entre chaque session.

Au sein d'un essai, la procédure était toujours la même. Les participants débutaient l'essai en cliquant sur un bouton en forme de triangle (bouton *play*) pour lancer la lecture de la vidéo cible. Cette vidéo était lue intégralement une seule fois (10 s) de manière isolée et sans distorsion à sa taille normale (160×120 px). La phase d'exploration débutait automatiquement dès la fin de la lecture de la vidéo cible. Au début de cette phase d'exploration, la collection de vidéo était présentée par une vue d'ensemble (facteur d'échelle égal à 1) de sorte que l'utilisateur puisse percevoir l'ensemble de la collection. Le participant pouvait ensuite utiliser les méthodes de zoom proposées et explorer la collection pour retrouver la vidéo cible.

### 4.2.8 Participants

Seize participants (11 hommes, 5 femmes) ayant des connaissances de base en informatique et familiers avec l'utilisation d'une souris ont participé à cette expérience. La moyenne d'âge

était de 27 ans. Les participants étaient rémunérés $CAD 15 pour leur participation.

## 4.3 Résultats et analyse

Les variables dépendantes incluaient des mesures de performance temps de réponse (RT) [3] et nombre d'erreurs, ainsi que des variables liées au ressenti utilisateur constituées de notes subjectives recueillies après chaque bloc sur l'adaptabilité, la difficulté et l'efficacité perçues, et de notes de préférence globale et de commentaires libres recueillis à la fin de l'expérience. Pour présenter les différents résultats sur les graphiques (figure 4.13, 4.14 et 4.15), nous utilisons un code couleur : les résultats de la méthode PZ sont représentés en vert tandis que les résultats de la méthode FL sont en bleu ; les conditions Visuel-seul (V) sont claires tandis que les conditions AudioVisuelles (AV) sont foncées.

Pour l'analyse statistique nous avons d'abord enlevé les essais incorrects pour lesquels une mauvaise vidéo a été sélectionnée. Cela représente environ 3,2 % des essais parmi les 960 essais : 5 erreurs pour AV-PZ, 2 pour V-PZ, 14 pour AV-FL et 10 pour V-FL (240 essais par condition). Ensuite pour l'analyse des RT, nous n'avons considéré que les essais corrects. Nous avons supprimé les valeurs aberrantes (*outliers*) pour chaque condition et participant (environ 6.5 % des essais, au total : 13 valeurs aberrantes pour AV-PZ, 11 pour V-PZ, 6 pour AV-FL et 7 pour V-FL), en considérant comme valeurs aberrantes tout RT supérieur à la moyenne des RT des réponses correctes plus deux fois la valeur de l'écart-type.

Une ANOVA croisée (2 modalités × 2 modes) a révélé que les RT sont significativement plus courts pour PZ que pour FL ($F(1,890) = 8.82$, $p = 0.003$) (voir figure 4.13). Aucun effet d'interaction entre les méthodes et les modalités n'a été observé ($F(1,888) = 0.06$, $p = 0.81$). Nous avons ensuite analysé séparément les méthodes FL et PZ pour comparer les rendus bimodaux AV et les rendus unimodaux V.

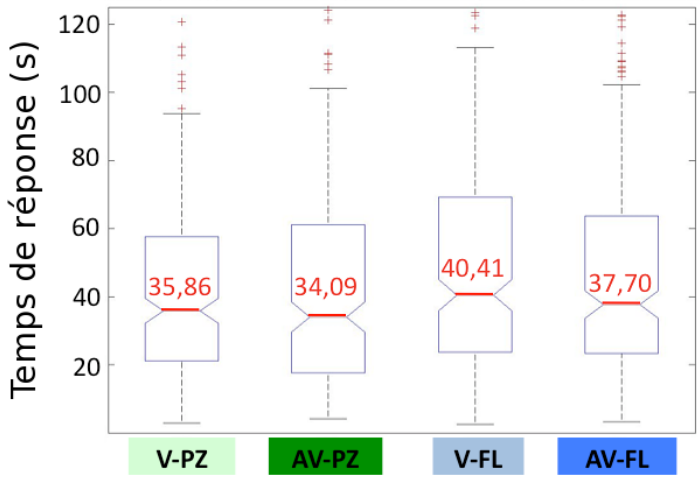



**Figure 4.13.** Temps de réponse calculés sur l'ensemble des participants et présentés sous forme de boîte-à-moustaches. La médiane et sa valeur sont présentées en rouge.

3. Les temps de réponse ici correspondent à des temps d'exécution.

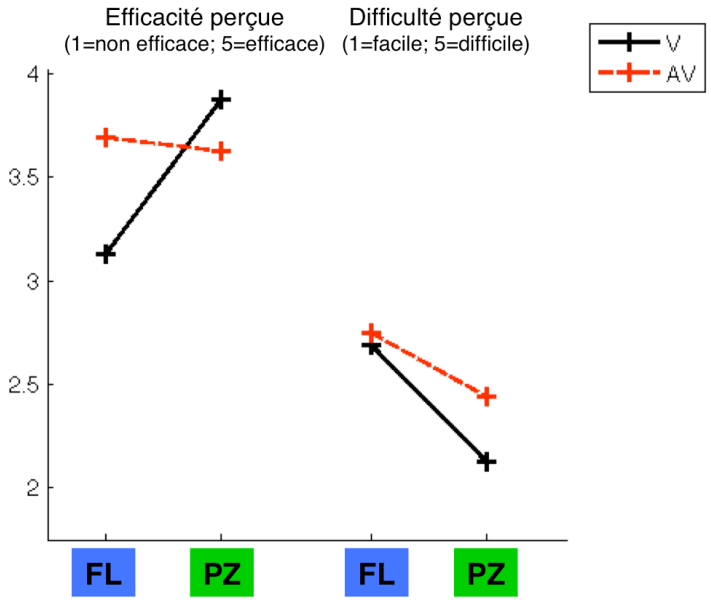



**Figure 4.14.** Notation moyenne sur l'ensemble des participants de l'efficacité et de la difficulté perçues.

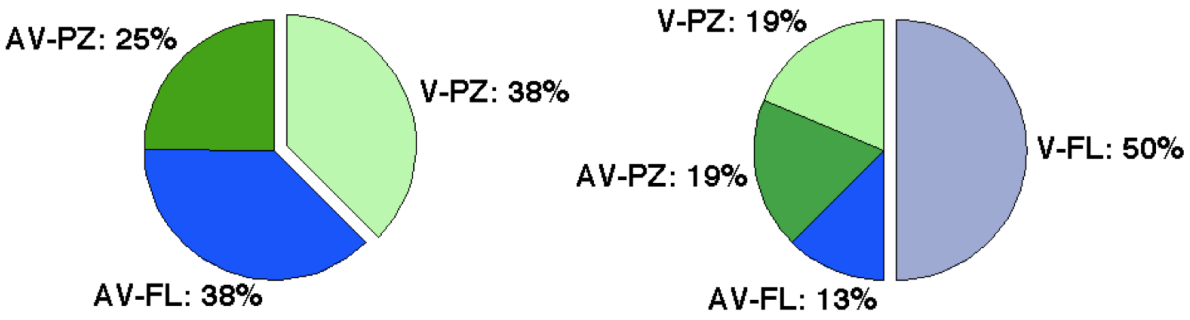



**Figure 4.15.** Répartition de la préférence globale des participants avec à gauche « la condition la plus appréciée » et à droite « la condition la moins appréciée ».

Pour la méthode PZ, aucun effet significatif n'a été observé entre les deux modalités ($F(1,448) = 0.46$, $p = 0.59$). Cependant, l'analyse des évaluations subjectives (figures 4.14 et 4.15) et des commentaires libres a indiqué que les participants ont perçu l'ajout d'audio négativement pour la méthode PZ. En effet, V-PZ a été significativement évaluée comme plus facile à utiliser que AV-PZ ($t(15) = 2.07$, $p = 0.05$, t-test). V-PZ a également été perçue comme plus efficace que AV-PZ, mais cette différence n'a pas atteint une significativité statistique. En commentaire libre, les participants ont indiqué que la méthode PZ « produit trop de bruits qui se chevauchent lors de la présentation de nombreuses vidéos »[4] ce qui est « plus une distraction qu'une aide »[5]. Ils ont aussi rapporté avoir des difficultés à associer chaque son à

4. "*produced too much overlapping noise when scanning many videos*"
5. "*more a distraction than an aid*"

la bonne vidéo car les sons présentés simultanément étaient trop nombreux. Ainsi d'après les participants, même dans le cas audiovisuel, la méthode PZ était utilisée « la plupart du temps comme un balayage visuel plutôt qu'audiovisuel »[6], d'autant plus que l'information visuelle était très fiable avec cette technique sans distorsion. Les participants ont aussi commenté avoir apprécié le rendu visuel offrant un « balayage visuel de nombreux objets de la même taille »[7] avec « une uniformité lors du défilement »[8].

Les RT pour la condition AV-FL sont inférieurs à ceux de V-FL (37.70 s contre 40.41 s), mais comme pour PZ, cette différence n'est pas significative ($F(1,442) = 0.06$, $p = 0.80$). Cependant, les résultats qualitatifs (figures 4.14 et4.15) et les observations faites par les participants diffèrent de ceux de PZ car, pour FL, la modalité n'affecte pas la difficulté mais l'audio améliore l'efficacité perçue ($t(15) = 2.52$, $p = 0.02$, t-test). Les participants ont justifié cela par le fait que le nombre de sons présentés simultanément était trop grand avec FL, mais moins qu'avec PZ. De plus ils ont ajouté que ce bruit était bénéfique pour un facteur d'échelle grand (« vue zoomée ») : « Au début, l'audio semble être un distracteur, mais une fois que la loupe est agrandie, il est utile »[9]. Ainsi, la lentille a permis aux participants de « balayer visuellement plusieurs vidéos mais seulement quelques unes en audio »[10]. Dans ce cas, la modalité auditive a alors été utilisée pour compenser le manque d'information fournie par la modalité visuelle que la vue en œil-de-poisson limite.

Pour résumer, en termes de techniques, PZ était plus rapide que FL et en termes de modalité, l'ajout d'un rendu audio avait un effet positif pour la méthode FL mais un effet négatif pour la méthode PZ. Ainsi, pour reprendre le commentaire d'un participant, « avec la méthode PZ, il y a trop de bruit, mais avec la méthode FL, c'est amusant puisque tu n'entends que 3 ou 4 sons. Mais c'est plus facile avec PZ parce que tu peux voir toutes les vidéos [11] ». Dans les commentaires libres, 81 % des 16 participants reportent se fier au rendu audio durant l'expérience, soit pour survoler la collection par le son (55 %), soit pour confirmer la sélection faite visuellement lorsque les vidéos sont ambigües (25 %). La figure 4.15 représente les préférences globales, c'est-à-dire les conditions préférées et les moins appréciées des participants. La majorité des participants (63 %) a préféré une condition bimodale, que ce soit AV-PZ ou AV-FL. En parallèle, un pourcentage comparable de participants (69 %) a désigné comme condition la moins appréciée une condition liée à une présentation unimodale. Ces deux dernières observations indiquent que l'ajout d'un rendu audio améliore le ressenti de l'utilisateur. Il reste à améliorer les métaphores et les réglages de chaque modalité, en particulier les réglages de l'audio, pour avoir un apport notable de la modalité auditive sur les performances des utilisateurs de cette interface.

6. "*mostly a visual scan instead of audio-visual*"
7. "*visual scanning of many items of the same size*"
8. "*the uniformity when scrolling* "
9. "*At first, the audio seemed to be a distracter, but once the magnifier is zoomed in, it is helpful* "
10. '*visually scan multiple videos while audio scanning just few*"
11. "*With PZ there are too much noise but with FL it's funny you've got only 3 or 4 sounds. But it is easier with PZ cause you can see all the videos*"

## 4.4 Discussion

Cette étude avait pour but de vérifier si l'ajout d'un rendu audio pouvait améliorer la navigation dans une collection de documents audiovisuels en utilisant une interface multimodale. Dans la première étape de ce travail, nous avons proposé plusieurs manières de combiner un rendu audio aux rendus graphiques pré-éxistants. Deux méthodes audiovisuelles, Pan&Zoom et lentille Fisheye, ont été implémentées dans une modalité visuelle seule et une modalité audiovisuelle. La comparaison de ces deux modalités a permis de mesurer la contribution de l'audio dans le rendu audiovisuel.

Aucune différence significative entre les RT des deux types de modalités, audiovisuelle multimodale ou purement visuelle, n'a pu être observée. Cela peut être expliqué par la prédominance de la vision sur ce type de tâche de recherche spatiale. En effet, comme on a pu le voir dans le chapitre 3 sur l'analyse des travaux en perception, la perception visuelle est très efficace pour détecter une cible parmi plusieurs distracteurs visuels. En revanche, l'audition est limitée sur des tâches où les sources sont entendues en concurrence. De plus, au-delà du fait que la vision semble plus adaptée pour réaliser ce type de tâche, les participants font remarquer qu'ils sont familiers avec les interfaces visuelles alors qu'ils utilisent rarement des interfaces sonores pour la navigation. Le manque de familiarité peut avoir nui à leur utilisation de l'audio dans le sens où les participants n'ont pas appris à extraire des informations auditives et à s'y fier dans ce type d'application de recherche de vidéos. Nous soulignerons donc que l'absence de différence significative indique également que, malgré la nouveauté de l'interaction proposée, la présence d'audio n'a pas dégradé les performances, ce qui est très encourageant.

D'ailleurs, en dépit de l'absence de gain de temps, les sujets ont rapporté que l'audio était un attrait amusant de l'interface et que c'était un moyen intéressant d'apporter de l'information supplémentaire. Au delà de la non familiarité des tâches de recherche sonore, nous pensons que si les performances ne sont pas améliorées en présence d'audio c'est qu'il y a une compensation entre les effets positifs (apport d'informations complémentaires ou redondantes) et des effets négatifs (fatigue auditive, difficulté à séparer les sources, gêne auditive). En effet, le bruit de fond créé par la lecture des sons des vidéos contextuelles à l'extérieur de la lentille, bien que de volume très faible, était noté comme *"ennuyant"*. Il est clair que la présence de sons sans intérêt direct pour l'utilisateur a augmenté la fatigue auditive. Pour le développement de futures interfaces, nous retenons qu'il aurait été plus efficace de rendre silencieuses toutes les sources à l'extérieur de la lentille, ou d'éteindre les sources non visibles dans la fenêtre graphique pour la méthode PZ, comme c'était le cas dans les interfaces sonores *SonicBrowser* [Fernström et Brazil 2001] et *SoundTorch* [Heise *et al.* 2008].

Par ailleurs, les participants ont reporté s'être fiés plus au visuel pour la navigation, surtout dans le cas où le rendu graphique ne présentait pas de distorsion, et où plsieurs vidéos étaient affichées simultanément avec la même taille (rendu PZ). Les notes des participants ainsi que leurs commentaires montrent en revanche que l'apport d'information par la modalité auditive est bénéfique surtout quand peu de sources sont entendues comme c'était le cas dans le rendu de type Fisheye. D'après ces résultats, on peut supposer qu'une méthode de rendu combinant les avantages des deux méthodes proposées dans cette étude sera plus

avantageuse que chacune des deux méthodes prises séparément. Ainsi une autre combinaison audiovisuelle, non testée dans cette étude, alliant une présentation Pan&Zoom pour le visuel et une présentation B+T plus sélective pour l'audio (comme dans le mode FL testé ici), aurait pu être plus efficace et plus appréciée des utilisateurs. Il faudrait réaliser une nouvelle évaluation pour comparer cette nouvelle combinaison aux méthodes précédemment testées.

Notre but premier n'était pas de comparer les techniques PZ et FL, mais bien d'évaluer la contribution de la modalité auditive pour chacune des techniques. Cependant les performances avec la méthode PZ ont été meilleures qu'avec la méthode FL. Bien que le même mode de contrôle ait été utilisé pour les deux méthodes, il est possible que cela ait introduit un biais dû à la difficulté de pointage. En effet, pour sélectionner une vidéo, le participant devait placer son curseur sur cette vidéo avant de cliquer. Pour réussir cette tâche plus facilement la taille de la vidéo devait être amplifiée par la technique de zoom disponible, FL ou PZ selon les cas. Or agrandir une vidéo avec la méthode FL nécessite de placer la lentille de façon quasi centrale au-dessus de la vidéo. Ce problème n'avait pas lieu avec la méthode PZ puisque avec cette méthode toutes les vidéos affichées ont la même taille quelle que soit leur position à l'écran. L'avantage de la méthode PZ pourrait donc être attribué à un problème de contrôle d'interaction. Ce biais devrait donc disparaître en appliquant de nouveau le mode de contrôle standard de chaque méthode : centre de la lentille attachée au curseur de la souris pour FL et, par exemple, barre de défilement pour PZ.

Dans l'évaluation menée, les vidéos étaient alignées et seul l'azimut était exploité pour la spatialisation des sources sonores. Un arrangement des vidéos sous forme de grille peut toutefois être utilisé, à condition de prendre en compte l'élévation des sources sonores pour leur spatialisation. Notre proposition de rendu audio de type Bifocal+Transparence peut être aisément généralisée à l'élévation, en reprenant par exemple les équations utilisées pour l'azimut et en les appliquant directement, ou avec de légères modifications empiriques, à l'élévation. Les seules contraintes sont alors les limites de la perception auditive et celles des techniques de restitution en matière d'élévation. Finalement, la méthode proposée ici dans un cadre d'exploration de collections vidéo pourra être aisément adaptée et réutilisée pour l'exploration d'environnements virtuels immersifs.

## 4.5 Conclusion

Dans ce chapitre nous avons présenté l'extension de méthodes audio et visuelles à l'audiovisuel pour la présentation de documents vidéos. Une méthode Pan&Zoom et une lentille grossissante audiovisuelle ont ainsi été implémentées en combinant une distorsion de la taille des objets en visuel et une distorsion de la position spatiale et du volume sonore en audio. Une étude d'utilisabilité a permis de tester ces deux méthodes sur une tâche appliquée de recherche de vidéo et d'évaluer la contribution de l'audio dans chacun des cas.

Les réactions positives des participants durant l'expérience ont montré le potentiel d'un rendu audio lorsque celui-ci concentre l'attention sur une petite quantité de sources sonores (3 ou 4 sources à la fois). Cette étude a donc confirmé que l'usage de l'audio est intéressant pour de la recherche de vidéos musicales. Ces résultats sont encourageants et nous invitent donc à poursuivre nos recherches sur la conception de stratégies audio ou audiovisuelles de documents multimédia.

D'autant plus que les résultats de cette étude confirment qu'il est possible de combiner des stratégies de présentation multi-échelles auditives et visuelles pour pallier le problème d'espace écran tout en rendant disponible le contenu sonore. Dans cette étude, nous avons utilisé des stratégies de présentation déjà exploitées dans les interfaces zoomables unimodales, visuelles ou sonores. Pour l'audio, nous avons exploité des distorsions de volume sonore sur lesquelles reposent la majorité des stratégies de présentation sonores. Or ces distorsions de volume pourraient être combinées à d'autres types de distorsion permettant elles aussi de mettre certaines sources sonores plus en avant. Nous présenterons, dans la suite du manuscrit (chapitre 6), une étude menée pour définir de nouvelles stratégies de présentation audio, développées à partir de paramètres acoustiques pas encore exploités dans les interfaces zoomables. En particulier, cette étude s'est concentrée sur les paramètres liés à la sensation de distance, comme la réverbération ou le rapport d'énergie entre les hautes fréquences et les basses fréquences.

Par ailleurs, cette étude apporte un élément de réponse à notre question de recherche sur l'optimisation de l'intégration des méthodes audio et visuelle et sur l'intérêt d'une distorsion cohérente entre les modalités. En effet dans ce cas précis, la distorsion visuelle semblait inutile voire désavantageuse tandis que la distorsion sonore était nécessaire. On peut en déduire qu'une combinaison audiovisuelle à base de distorsion ne nécessite pas de déformer de la même manière le rendu audio et le rendu visuel.

Finalement, selon les participants, le principal défaut de l'audio dans cette étude a été le problème de la fatigue auditive liée à la difficulté pour les participants de démêler les différents flux sonores présentés simultanément. Il s'agit d'un problème de perception auditive qu'il est nécessaire de mieux cerner et contourner pour développer de meilleures interfaces sonores. Dans la suite du manuscrit nous avons donc orienté nos recherches sur cette problématique de la perception sonore lors de la présentation de multiples sources concurrentes. Dans le chapitre 5 nous exposerons une étude où nous avons observé les difficultés d'identification des sons dans le bruit et évalué dans quelle mesure la combinaison avec un rendu visuel pouvait réduire ces difficultés. Dans le chapitre 6 nous essaierons au contraire de mettre à profit ces phénomènes perceptifs et attentionnels, tels que décrits dans le chapitre 3, pour créer de nouvelles stratégies de présentation auditive multi-échelles.

## Troisième partie

# Apports en perception multisensorielle : dégradations et distorsions multimodales

Chapitre 5

# Identification de sons en présentation simultanée et influence du contexte visuel : cas des sons d'environnement dans le bruit

*Images et sons comme des gens qui font connaissance en route et ne peuvent plus se séparer.*
Robert Bresson, *Notes sur le cinématographe*, 1995

## Sommaire

Ce chapitre présente une étude sur l'identification de sons d'environnement en condition d'écoute dégradée et s'interroge sur le rôle du contexte visuel pour la reconnaissance de sons. Bien que cette étude ne soit pas directement en rapport avec notre problématique de recherche sur les stratégies de présentation de documents multimédia, nous en rapportons les principaux résultats qui confirment l'avantage de la multimodalité et de la multisensorialité. Ces résultats complètent ainsi les travaux présentés au chapitre précédent.

Cette étude a été menée en collaboration avec Catherine Guastavino, Bruno Giordano et Ilja Frissen des laboratoires MIL (*Multimodal Interaction Laboratory*) et CIRMMT (*Centre for Interdisciplinary Research in Music Media and Technology*) à Montréal, Canada. Les résultats de cette étude ont fait l'objet d'une publication intitulée "*Effect of Signal-to-Noise Ratio and Visual Context on Environmental Sound Identification*" lors de la 128$^{\text{ème}}$ convention de l'*Audio Engineering Society* à Londres.

## 5.1 Introduction

Dans le but de développer des interfaces multimédia pour la recherche de documents vidéos, nous avons décidé de nous concentrer sur des méthodes où les documents sont présentés simultanément. De ce fait, il est important de comprendre comment sont perçues différentes images affichées en parallèle, mais aussi comment sont entendus différents sons présentés en même temps. Or, lorsque plusieurs sons sont entendus simultanément il en résulte une sorte de bruit de fond. La ségrégation entre les différentes sources sonores est alors plus difficile à mettre en place. De manière intuitive, nous savons qu'il est plus dur d'identifier un son si celui-ci est noyé dans le bruit. Certains sons sont alors masqués et parfois même indiscernables. Cela a été démontré pour la parole qui devient moins intelligible lorsque plusieurs locuteurs s'expriment simultanément ou lorsque l'environnement acoustique environnant est bruyant. Toutefois, malgré leur importance pour la compréhension de l'espace dans lequel nous sommes plongés, aucune étude sur l'identification des sons d'environnement dans le bruit n'a été effectuée auparavant. Ce sera l'objet de ce chapitre.

Par ailleurs, plusieurs études sur l'identification de la parole en condition dégradée ont démontré l'intérêt d'utiliser une présentation visuelle simultanée pour compenser la perte d'information auditive induite par le bruit. Nous nous proposons de tester cette compensation visuelle sur les sons d'environnement.

Nous cherchons donc ici à quantifier l'effet de la dégradation due à l'ajout de bruit, et à évaluer la compensation éventuelle par un contexte visuel sur l'identification des sons d'environnement. Nous avons séparé cette étude en deux phases : d'une part, la création d'un ensemble de stimuli de type sons d'environnement accompagnés d'une représentation visuelle et, d'autre part, une comparaison de plusieurs niveaux de bruits et plusieurs contextes visuels, congruents ou non, sur une tâche d'identification des sons.

## 5.2 Termes et définitions relatifs aux sons d'environnement

Les ***sons d'environnement***, aussi appelés *everyday sounds* en anglais, peuvent être définis comme les sons non relatifs à la parole, non musicaux et que l'on peut trouver dans des

environnements du quotidien. Ces sons jouent un rôle de première importance pour ce qui est de la prévention de dangers potentiels (alarmes) ou de la description d'un environnement et la sensation d'y être plongé (trafic et foule pour une environnement urbain par exemple, ou au contraire chants d'oiseaux...). Ils apportent des renseignements sur les objets et les évènements et sont, en général, perçus d'abord en termes de sources acoustiques. C'est seulement lorsque l'identification de la source d'origine est difficile que le son est décrit en fonction de ses qualités [Marcell *et al.* 2000].

### 5.2.1 Pourquoi s'intéresser aux sons d'environnement ?

Plusieurs raisons nous ont poussée à nous concentrer sur les sons d'environnement plutôt que sur d'autres types de sons.

Une des raisons initiales est le manque d'études menées sur les sons d'environnement. En effet, plusieurs travaux ont déjà été menés sur l'effet de la dégradation d'écoute pour la perception de la parole et de la musique, mais peu sur les sons d'environnement. Pourtant, comme l'indiquent les témoignages des personnes mal-entendantes [Gygi et Shafiro 2007b], ces sons sont primordiaux pour la compréhension de l'environnement et pour l'immersion. Ils sont donc très utilisés en réalité virtuelle, mais aussi au cinéma puisque les sons d'ambiance et les effets sonores permettent à la fois de renforcer l'information sur le lieu/la scène (ambiance de restaurant, bruits de circulation, chants d'oiseaux...) mais aussi de donner une sensation de présence des personnages (frottements de tissus, respiration...) [Chion 1990, p. 75].

De plus, les sons environnementaux proposent une large palette de significations (aspect sémantique) tout en étant de complexité spectro-temporelle de même ordre de grandeur que la parole, c'est-à-dire de grande complexité. En revanche, les sons environnementaux sont dépourvus de règles grammaticales ou lexicales propres au langage, règles qui influencent et facilitent la compréhension de la parole. Ainsi ces sons peuvent être une alternative à l'étude des sons de voix parlée car ils permettent de séparer les aspects linguistiques, cognitifs et psycho-acoustiques.

Indépendamment du cadre spécifique à la recherche de documents vidéos, en savoir plus sur l'écoute simultanée de plusieurs sons d'environnement peut être bénéfique pour d'autres interfaces audio, notamment :

- des interfaces d'**exploration de bases de données audio** spécifiques, dédiées aux ingénieurs et monteurs sons, permettant de chercher un son d'environnement particulier parmi une grande collection de sons d'environnement [Heise *et al.* 2008] ;
- des interfaces basées sur la **sonification** de données, c'est-à-dire la représentation par le son de données initialement non sonores (comme par exemple la présentation de molécules) [Hermann *et al.* 2011] ;
- tout type d'interface utilisant des **icônes sonores** (*auditory icons* [Gaver 1989; Brazil et Fernström 2011a] ou *earcons* [Brazil et Fernström 2011b; McGookin et Brewster 2004]) pour renseigner l'utilisateur sur l'état de certains processus (« allumage de l'ordinateur », « détection d'un périphérique », « vidage de la corbeille »,...) ;
- l'**exploration d'environnements virtuels** complexes qui nécessitent souvent une simplification du contenu sonore afin de le rendre audible tout en restant crédible [Moeck *et al.* 2007].

Gygi et Shafiro ont rendu disponible des revues de littérature sur l'utilisation des sons d'environnement en général [Gygi et Shafiro 2007b], et notamment dans les interfaces homme-machines [Gygi et Shafiro 2009].

### 5.2.2 Exemples de sons d'environnement et catégorisation

Une difficulté majeure dans l'étude des sons environnementaux vient de leur large nombre et de leur très grande diversité. L'industrie des *sound effects* est fructueuse et de nombreuses banques de sons sont disponibles dans le commerce sur CDs ou par site Internet. La diversité des sons se retrouve à la fois sur le plan sémantique (tout type d'environnements, d'objets...) et sur le plan acoustique (grande variabilité dans le spectre et la structure temporelle). De cette grande variabilité naît le besoin de classifier les sons.

Les banques de sons organisent les sons plutôt selon leur contexte et l'environnement dans lesquels on les trouve. Par exemple, on distinguera les « sons de forêt » (oiseaux, vent, scie, véhicules, rivière) des sons de « cuisine » (bouilloire, crépitement d'huile, téléphone) ou de « bureau » (téléphone, papier froissé, clavier), même si une source peut se retrouver dans plusieurs environnements distincts [Gaver 1993, p. 13]. Au contraire, les études scientifiques préfèrent organiser les sons sous forme hiérarchique en fonction du mode de production du son et non plus du contexte.

Gaver [Gaver 1993] est le premier à proposer une classification des sons sous cette forme (voir figure 5.1). Les sons étudiés sont classés en trois grandes catégories, sons de liquide, sons de solide en vibrations et sons aérodynamiques, en distinguant les sources en fonction des matériaux intervenant (liquide, solide ou gaz). Ces catégories sont ensuite séparées en sous-catégories en fonction de l'interaction entre le ou les différent(s) matériau(x), par exemple : impact ou frottement pour les sons de solides.

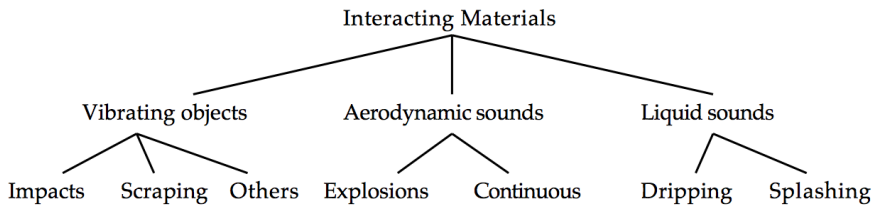



**Figure 5.1.** Classification hiérarchique selon [Gaver 1993]. Les sons, de sources non vivantes, y sont classés selon le matériau et le mode d'interaction.

Cette catégorisation ne fait intervenir que des sons d'objets ou de choses non vivantes. Au contraire, Lewicki [Lewicki 2002] propose de grouper les sons environnementaux selon 2 catégories seulement : « vocalisations animales » et « tout le reste ». Selon lui, le codage neuronal mis en place par le système auditif pour ces 2 classes de sons est distinct. Cette distinction est confirmée dans d'autres études comme [Lewis *et al.* 2005].

Les études plus tardives sur les sons d'environnement reprennent en partie la classification de Gaver, en tenant compte aussi des sons d'animaux. Par exemple Gygi et ses collègues [Gygi *et al.* 2004 2007] différencient les classes : sons humains non verbaux / sons d'animaux (ou

vocalises animales) / sons de machines / sons relatifs à des conditions météorologiques / sons générés par des activités humaines. Shafiro et ses collègues [Shafiro 2008ba] regroupent les sons d'humains ou d'animaux en une seule classe. Ils reprennent les catégories de Gaver pour les sons de sources non vivantes, en ajoutant une classe pour les « sons électriques / électroniques ou de signalisation ». Cette même distinction pour les sources non vivantes se retrouve dans [Ozcan et van Egmond 2009].

La classification que nous retiendrons est une synthèse des différentes catégories exposées ci-dessus. Elle est présentée en figure 5.2. La distinction pour les sons de sources vivantes en sous-catégories vocalisation / locomotion / alimentation reprend les spécifications des stimuli étudiés dans [Giordano *et al.* 2010] que nous utiliserons par la suite dans la phase expérimentale de ce chapitre.

### 5.2.3 Facteurs pour la reconnaissance de sons d'environnement

Plusieurs études se sont intéressées à l'identification des sons d'environnement et aux facteurs qui permettent cette identification. Pour plus de détails, le lecteur pourra se référer aux revues de littérature [Gygi 2001; Gygi et Shafiro 2007b]. Ces études ont montré qu'il était possible pour l'être humain de discerner la forme, la taille ou encore le matériau d'un objet, simplement au son qu'il provoque quand on le frappe sur un autre. D'autres études ont montré qu'il était même possible de définir le genre d'une personne en train de marcher simplement en entendant le bruit de ses pas [Li et Pastore 1991]. Les facteurs peuvent être acoustiques et dépendre de la source elle-même et du système auditif de l'auditeur, ou plutôt cognitifs et dépendre de l'auditeur uniquement, de ses attentes et de ses connaissances préalables.

#### 5.2.3.1 Facteurs acoustiques

Les paramètres acoustiques qui permettent d'identifier et de catégoriser les sons d'environnement sont mis en évidence à travers des études perceptives, mais aussi à travers des études sur l'analyse et la classification automatique de sons. Ces études reprennent en particulier les notions de descripteurs utilisés en reconnaissance de la parole ou en indexation. D'ailleurs, comme pour la reconnaissance de la parole, l'identification des sons d'environnement repose sur des facteurs spectro-temporels [Gygi *et al.* 2004].

Les **variations temporelles** permettent à elles seules, c'est-à-dire sans information fréquentielle disponible, l'identification de 50 % des sons avec une exactitude d'au moins 50 % [Gygi *et al.* 2004, Expérience 3]. Ces résultats sont obtenus à partir de sons modulés par du bruit (*Event Modulated Noise* ou EMN). Dans un EMN, l'enveloppe temporelle du signal d'origine est extraite par filtrage puis excitée par du bruit blanc large bande. Le signal obtenu possède donc la même information temporelle que le signal d'origine presque plus d'information spectrale car le spectre est quasiment plat (toutes les fréquences ont la même énergie). Dans cette étude, les auteurs remarquent que l'analyse temporelle permet d'autant mieux la reconnaissance d'un son que ses variations temporelles (pattern temporel) sont très marquées et que le contenu fréquentiel original est déjà très étalé. Ainsi l'identification des sons d'impact, avec un contenu spectral proche du bruit et très courts temporellement, de même que celle des sons rythmés,

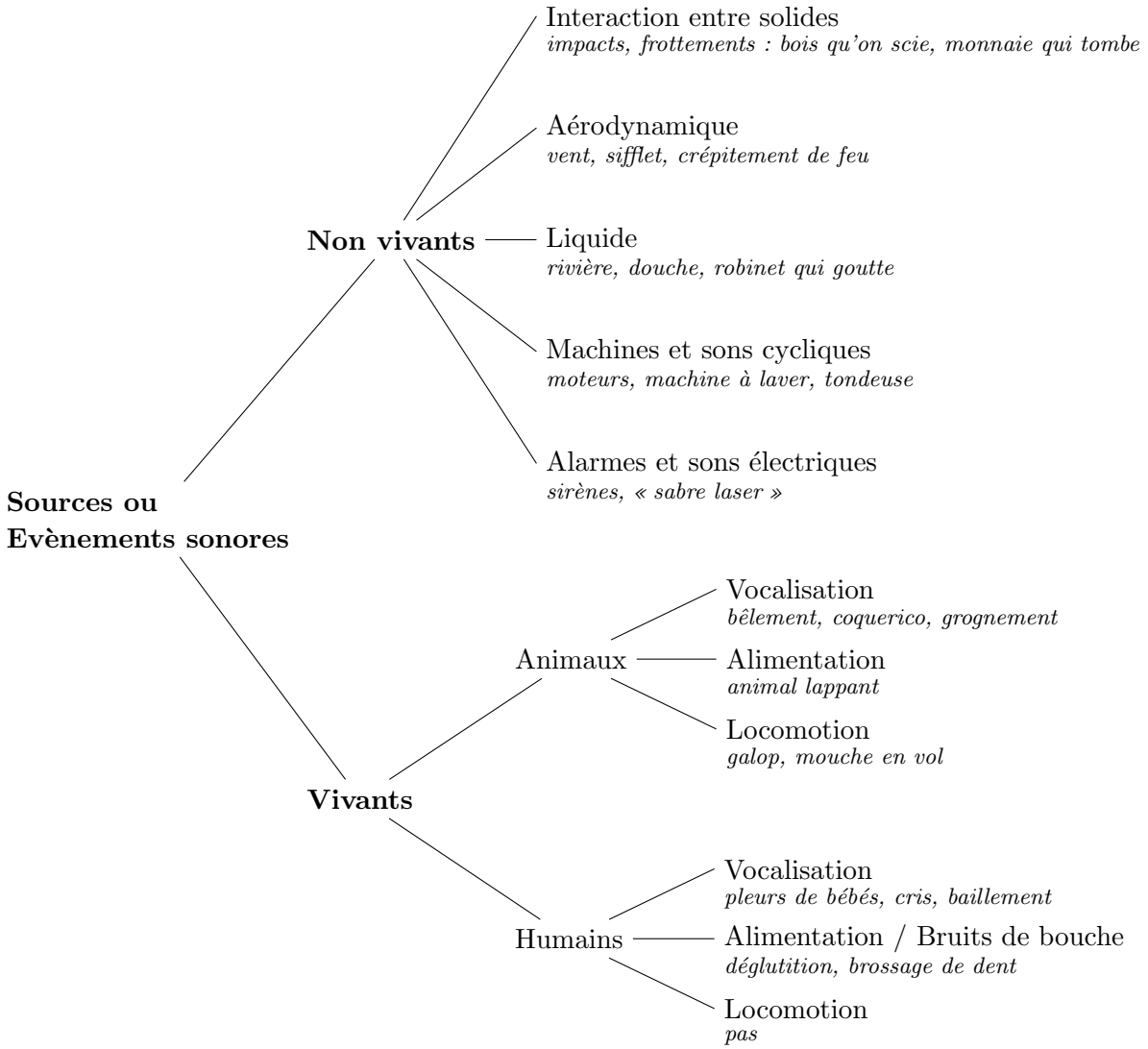



**Figure 5.2.** Différentes catégories de sons d'environnement. Des exemples d'évènements ou sources sonores sont présentés en italique.

par exemple le « galop d'un cheval », repose en grande partie sur l'aspect temporel.

Le **contenu fréquentiel** est lui aussi essentiel, en particulier pour différencier deux exemplaires d'un même son (comme deux aboiements différents [Riede *et al.* 2001]). Dans [Gygi *et al.* 2004, Expériences 1 et 2], les auteurs filtrent les sons selon différentes bandes fréquentielles et mettent ainsi en avant l'importance de l'information spectrale dans l'identification [Shafiro 2008b]. En particulier certaines zones du spectre (entre 1200 Hz et 2400 Hz) apportent plus d'informations que les autres.

Les paramètres acoustiques relatifs au contenu spectral peuvent être mesurés à travers différents indices acoustiques et leurs variations au cours du temps, comme des mesures de l'étalement spectral et du degré d'harmonicité des sons [Gygi *et al.* 2004 2007; Misdariis *et al.* 2010].

Le paramètre de **centroïde spectrale** donne une indication globale sur la répartition fréquentielle. Ce paramètre a une forte connexion avec la sensation de brillance d'un son. Il indique la moyenne des fréquences présentes dans le signal sonore, pondérée par l'amplitude de chaque fréquence. Quatre études de catégorisation libre comparées dans [Misdariis *et al.* 2010] montrent qu'il s'agit d'un paramètre principal d'après l'analyse en composantes principales.

Le paramètre d'**harmonicité** est aussi extrêmement important et revient comme facteur principal dans trois des études décrites par [Misdariis *et al.* 2010]. Chaque signal sonore est composé de deux composantes, l'une harmonique (composante périodique), l'autre bruitée (non harmonique). L'harmonicité d'un signal est définie par le rapport d'énergie entre ces deux composantes. On parle de *Harmonic-to-Noise ratio* (HNR). Plus un son est harmonique et plus son HNR augmente ; lorsque le niveau de la composante bruitée est égal au niveau de la composante harmonique, le HNR est nul. Ce paramètre a d'abord été défini comme indice du degré d'enrouement d'une voix (plus une voix est enrouée moins elle est harmonique) puis pour décrire la qualité d'une voix en général [Yumoto *et al.* 1982]. Plusieurs études ont ensuite démontré qu'il s'agit d'un paramètre acoustique central pour l'identification et la description de sons d'environnement [Riede *et al.* 2001; Gygi et Shafiro 2007b; Gygi *et al.* 2007]. En particulier, il semble que ce paramètre permette surtout de distinguer deux classes : a) les sons d'alerte (alarmes, cloches) ou produits par des animaux (dont la majorité des sons de vocalisations harmoniques [Riede *et al.* 2001]) et b) les sons produits par des sources non vivantes (comme des sons d'impact ou reliés à l'eau) qui sont principalement inharmoniques et présents en *quasi* continu dans notre environnement, donc de moindre intérêt d'après la théorie de l'information.

#### 5.2.3.2 Facteurs cognitifs

D'après McAdams [McAdams 1993], lorsque des sons à comparer sont issus de catégories trop distinctes, ce ne sont plus les facteurs psychoacoustiques qui sont pris en compte mais des facteurs cognitifs de plus haut niveau. Cela sous-entend que la catégorisation se fait à partir de critères cognitifs (mais impliquent quand même la perception d'invariants acoustiques au sein d'une même catégorie) mais que la comparaison de deux sons issus d'une même catégorie demande un traitement spécifique et plus fin des paramètres acoustiques qui en ont permis la catégorisation.

Les nombreuses études de Ballas et ses collègues ont permis de relever plusieurs facteurs psychoacoustiques, mais aussi cognitifs. Par exemple dans [Ballas 1993] sont exposés les facteurs suivants :

L'**incertitude causale** : Plus le nombre de causes (sources/évènements) auxquelles un son peut être attribué est grand et plus l'identification de ce son sera longue et imprécise. L'incertitude causale est donc une mesure de l'ambiguïté d'un son. Elle se mesure sous forme d'entropie causale, ainsi la confiance dans l'identification d'un son diminue lorsque le nombre de causes augmente. Dans l'étude que nous avons menée et qui est détaillée dans la suite de ce chapitre, nous avons veillé à choisir des sons non ambigus (à incertitude causale *quasi* nulle).

La **familiarité** : Plus les participants sont familiers avec l'évènement ou la source ayant produit un son, selon des mesures obtenues par des échelles subjectives, plus les sons sont identifiés rapidement. De plus, les sons les plus familiers sont aussi ceux qui sont nommés le plus fidèlement (avec le plus de détails) et nommés par des termes communs au plus grand nombre d'auditeurs [Marcell *et al.* 2000].

La **typicalité** : [Ballas 1993] et [Guillaume *et al.* 2004] ont montré que plus un son est typique de sa catégorie de sons, plus il est identifié correctement et rapidement. En effet, un même type d'évènement sonore peut conduire à des sons plus ou moins similaires les uns des autres. Tous ces sons forment un exemplaire d'une même catégorie. Par exemple, les sons de klaxon peuvent être très différents d'un véhicule à l'autre, pourtant ils sont tous identifiés sous le terme « klaxon » grâce à plusieurs paramètres acoustiques typiques.

L'**attention** : Un son peut-être simplement entendu, c'est-à-dire perçu sans qu'on cherche réellement à l'identifier ou même à noter sa présence, ou au contraire écouté. On peut alors volontairement porter son attention sur ce son, soit pour en identifier la source ([Gaver 1993] parle alors d'« écoute quotidienne »), soit pour analyser ses attributs acoustiques (« écoute musicale »). De plus la capacité d'attention sélective en audition permet de se concentrer sur certains paramètres particuliers du son, on peut par exemple se focaliser sur une zone fréquentielle donnée. Cette sélection influence alors l'identification du son car elle met en avant certaines caractéristiques au lieu. De plus, il est possible qu'on s'attende à certains sons compte tenu de l'environnement dans lequel on se trouve (notion de contexte). Un son improbable pourra alors sauter aux oreilles [Gygi et Shafiro 2011] et requérir une identification plus rapide. C'est le cas d'une porte qui claque par exemple.

### 5.2.4 Dégradations de la condition d'écoute

Nous venons de présenter plusieurs facteurs cognitifs qui influencent l'identification de sons issus de la même catégorie. Deux autres facteurs, détaillés dans les paragraphes suivants, sont liés non plus au son à identifier en lui-même mais à son environnement, notamment quand ce dernier est bruyant et donc que la condition d'écoute est dégradée. La **connaissance préalable de l'environnement**, ou **contexte**, et l'influence des **autres modalités sensorielles** sont alors deux autres facteurs cognitifs qui vont influencer l'identification.

#### 5.2.4.1 Cas de la parole

La recherche sur la perception et l'intelligibilité de la parole (compréhension des mots ou de la phrase employés) en condition dégradée est extrêment riche. Les dégradations étudiées sont diverses, allant de l'ajout de bruit de fond à la diminution d'informations spectro-temporelles, comme le filtrage fréquentiel pour la simulation de problèmes d'audition. Ces études permettent d'évaluer les capacités humaines à récupérer l'information utile pour la compréhension de la parole, malgré les difficultés dues à la dégradation du signal ou de l'environnement.

Nous nous intéresserons essentiellement ici à la perception de la parole dans le bruit. Cette thématique s'apparente aux notions de masquage et au problème *cocktail party* dans lequel il s'agit d'écouter et identifier la parole d'un locuteur parmi plusieurs (voir section 3.3.2.1, p. 37 et [Bronkhorst 2000]). Des études sur l'intelligibilité de la parole dans le bruit (*speech-in-noise*), nous retiendrons plusieurs facteurs d'influence. Le premier correspond au rapport de niveau sonore entre celui de la parole à identifier et celui du bruit de fond (*target-to-masker ratio*), encore appelé ***rapport signal-sur-bruit*** (**RSB** ou *signal-to-noise ratio*). Ce facteur est assez intuitif : plus le fond sonore est fort en volume sonore et plus l'identification de la parole est difficile. Toutefois, des études ont montré que pour un même niveau de bruit de fond, le **type de bruit de fond** (ou type de masqueurs) pourra améliorer l'identification ou la rendre encore plus difficile. Ainsi la difficulté sera différente pour un bruit stationnaire (bruit blanc non modulé) comparé à un bruit fluctuant, c'est-à-dire avec une amplitude variable au cours du temps (un bruit modulé, une ambiance...) [Howard-Jones et Rosen 1993]. De plus, la difficulté d'identification est augmentée lorsque le fond sonore est aussi apparenté à de la parole (condition avec plusieurs locuteurs, simulation par *babble-noise*, etc...) [Hawley *et al.* 2004].

Par ailleurs, comme nous l'avons vu section 3.4.1.2, p. 49, la façon dont l'interaction audiovisuelle influence la perception de la parole fait l'objet de nombreuses études. Il est alors maintenant acquis que la présence d'indices visuels de la parole permet de compenser, au moins partiellement, la dégradation due à un bruit de fond par rapport à une présentation sonore seule. Les premières études sur le sujet ont montré que l'apport du visuel, c'est-à-dire le gain de la condition multimodale par rapport à la condition audio seul, est d'autant plus grand que le niveau de bruit est fort et que l'identification de la parole sans indice visuel est faible (taux de réponses correctes ou %Correct faible) [Sumby et Pollack 1954]. Ces observations viennent étayer la théorie de l'efficacité inverse de l'intégration audiovisuelle (voir p. 54). Cependant des études plus tardives montrent que l'apport du visuel est plus fort pour un RSB intermédiaire ($\sim -12$ dB) et non pour un RSB plus faible [Ross *et al.* 2007; Ma *et al.* 2009]. Les résultats de [Ma *et al.* 2009] sont présentés figure 5.3.

#### 5.2.4.2 Cas des sons d'environnemment

Bien que leur nombre augmente, peu d'études ont été menées auparavant pour mesurer l'effet des dégradations sur la perception des sons d'environnement. Ainsi les principales études concernaient les effets des implants cochléaires pour mal-entendants, en particulier les effets du filtrage fréquentiel et d'une baisse de la résolution fréquentielle [Shafiro 2008ba]. Plus le nombre de canaux fréquentiels augmente et plus l'identification des sons est correcte.

Toutefois, nous n'avons trouvé aucune étude préalable concernant l'effet du masquage. Les seules études que nous avons pu trouver à ce sujet concernent plutôt l'influence d'un contexte

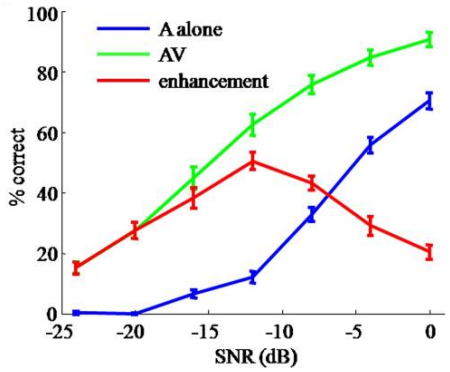



**Figure 5.3.** Données obtenues (moyennes + erreur standard) dans l'étude de [Ma *et al.* 2009] pour une condition audio seul (bleu) et audiovisuel (vert). L'apport du visuel (rouge) représente la différence entre la condition audiovisuel et la condition audio seul.

sonore informatif (ayant un lien sémantique) [Gygi et Shafiro 2011]. Dans ce type d'étude le fond sonore masque partiellement la source et le rapport entre le niveau de la source à identifier et celui du fond a de l'influence sur le taux d'identification.

L'étude que nous présentons dans la suite de ce chapitre a pour but de vérifier si la présence d'un bruit de fond provoque les mêmes effets négatifs sur la perception des sons d'environnement que sur la perception de la parole.

### 5.2.5 Influence du contexte

Dans la vie courante, les sons arrivent rarement isolés. La perception auditive de ces différents sons forme une scène auditive, et l'identification d'un seul son est influencé par la présence d'autres sons. De plus, chaque son perçu est issu d'un environnement complexe, constitué de plusieurs objets que l'on perçoit à travers plusieurs modalités sensorielles, qui fournissent un contexte au son.

Pour un environnement donné, certains sons « collent » mieux que d'autres à l'environnement. Ainsi la probabilité que ces sons apparaissent dans cet environnement est grande. On peut considérer ces sons comme ***congruents*** tandis que d'autres sons, peu probables dans cet environnement, seront considérés comme ***incongruents***. Par exemple, et pour reprendre un exemple de [Gygi et Shafiro 2007a], des « *sons de couverts et d'assiettes* » sont courants dans un restaurant, tandis que le « *hennissement d'un cheval* » ne l'est pas.

Il est communément admis que l'identification des sons en général est plus difficile en absence de contexte [Mynatt 1994]. Par exemple [Bilger *et al.* 0001] montre qu'un contexte sonore aide à percevoir la parole dans le bruit. Bien que peu nombreuses, quelques études se sont également intéressées à l'influence du contexte sur la perception des sons d'environnement. Ces études considèrent le contexte sous différentes formes. Il peut en effet donner plus d'informations concernant la source elle-même, sous une autre modalité sensorielle par exemple, ou au contraire sur l'environnement et les différents objets qui le composent. Dans ce dernier cas, le contexte pourra alors être sonore (sous forme d'ambiance sonore par exemple) ou visuel. Enfin, le contexte peut être présenté simultanément au son à identifier ou alors avant que le son n'ait lieu. Dans ce dernier cas, le contexte est appelé amorce.

#### 5.2.5.1 Contexte sonore en amorce

Ballas et Mullins [Ballas et Mullins 1991] proposent d'étudier le contexte sonore sous forme d'une séquence de sons isolés. Dans cette étude, les sons à identifier sont en fait des couples de sons ambigus ($son_1$,$son_2$), tels que chaque $son_1$ est facilement confondu avec le $son_2$. Par exemple les sons de « friture » (*food frying*) sont souvent confondus avec ce son d'« une mèche de détonateur allumée » (*fuse burning*). Les participants doivent identifier le $son_1$ présenté individuellement au sein d'une séquence de sons contruite. Cette séquence peut former un scénario congruent si l'ensemble des sons de la séquence se rapporte à un environnement dont le $son_1$ pourrait être issu (exemple pour la friture : son de couteau (trancher puis couper de la nourriture) puis friture), ou au contraire un scénario incongruent lorsque les sons semblent provenir d'un environnement correspondant au $son_2$ (toujours sur le même exemple : son d'allumette puis de friture puis d'explosion). L'étude compare ces deux conditions à deux autres conditions « contrôle » : une condition aléatoire dans laquelle la séquence de sons est construite à partir de sons choisis aléatoirement (indiquant le $son_1$ et le $son_2$ à parts égales), et une condition sans contexte (la source est présentée seule). Les résultats de cette étude montrent que la présence d'un contexte congruent donne de meilleurs taux d'identification que dans les cas de contexte incongruent ou aléatoire. En revanche les taux d'identification ne sont pas améliorés par rapport à la condition sans contexte. Les auteurs en déduisaient que la présence d'un contexte sonore congruent permet seulement de compenser les effets négatifs de la présence d'autres sons [Ballas et Mullins 1991] :

> *the only positive effect of consistent contexts is to offset the negative effects of embedding sound in a series of other sounds.*

On peut, au moins partiellement, expliquer cela par la difficulté pour l'auditeur à identifier d'abord chaque son de la séquence avant de pouvoir les relier les uns aux autres. Comme l'identification n'est pas parfaite pour chacun des sons isolés de la séquence, le bénéfice du contexte est limité.

La séquence de sons isolés proposée par [Ballas et Mullins 1991] est une forme d'amorce puisque certains sons de la séquence sont présentés avant le son à identifier et qu'aucun des sons n'est présenté simultanément. Dans le même ordre d'idée, Gérard présente dans sa thèse [Gérard 2004] une analyse des effets de la durée et du type d'amorce. Les amorces utilisées, tantôt congruentes, tantôt incongruentes, se trouvent sous différentes formes sonores. Il peut s'agir d'un son d'environnement, d'une ambiance sonore (enregistrement d'une scène auditive naturelle), d'un mot, ou d'une phrase décrivant l'environnement. Ces amorces représentent dans certains cas la source elle-même, dans d'autres cas une autre source reliée sémantiquement (comme le « miaulement d'un chat » pour amorcer l'« aboiement d'un chien »). Les résultats obtenus diffèrent selon le type d'amorce. Pour des amorces du type sons isolés ou langage, les sons sont mieux identifiés et plus rapidement avec des contextes congruents, ce qui est cohérent avec les résultats de [Ballas et Mullins 1991]. En revanche, un résultat inattendu est observé pour des amorces de type ambiance sonore : les sons sont mieux identifiés lorsque le contexte est incongruent.

#### 5.2.5.2 Contexte sonore simultané

Cette notion d'ambiance sonore est étudiée également dans les travaux de Gygi et Shafiro [Gygi et Shafiro 2007a 2011]. Toutefois, au contraire des travaux de Gérard [Gérard 2004], l'ambiance sonore n'est plus utilisée comme amorce mais est superposée temporellement avec le son à identifier. La scène auditive est alors utilisée comme fond sonore. L'étude visait à évaluer l'importance du niveau sonore de la source par rapport au niveau sonore de la scène (le rapport *So/Sc*). La paire son+scène auditive pouvait là encore être congruente, incongruente ou neutre. Comme dans les études de Gérard, ces études révèlent ce que les auteurs appellent un ***avantage de l'incongruence*** (ou *Incongruency Advantage*) : l'ambiance incongruente donne de meilleurs taux d'identification ($\sim 5$ %) que l'ambiance congruente.

Toutefois, cet avantage n'a lieu que pour certains rapports *So/Sc*. En effet, lorsque le niveau sonore de la source à identifier est trop faible par rapport au fond sonore ($So/Sc < -7.5$ dB), cet avantage disparaît. Ces travaux prouvent que la congruence sémantique n'est pas toujours favorable. De plus, dans ce cas où le fond sonore apportait en lui-même des informations, l'augmentation du niveau de bruit par rapport au niveau du signal utile dégrade l'identification qui devient moins précise et plus lente. On peut se demander si ces effets seront encore observables si le fond sonore n'apporte pas d'information sémantique (bruit blanc) et si le contexte d'ambiance sonore est remplacé par une représentation visuelle de la source sonore.

#### 5.2.5.3 Contexte visuel

Nous avons vu à travers les études mentionnées au paragraphe 3.4.1.2.b) qu'une représentation visuelle peut influencer la perception et l'identification de sons non liés à la parole. Etant distincte du stimulus sonore à identifier, cette représentation visuelle peut donc être considérée comme du contexte.

Comme pour le contexte sonore [Ballas et Mullins 1991; Gérard 2004], le contexte visuel peut-être présenté en amorce, avant ou après le son à identifier [Schneider *et al.* 2008]. Il peut aussi être simultané et présenter l'environnement où se situe la source à identifier (lien indirect) [Ozcan et van Egmond 2009], ou représenter la source elle-même [Suied *et al.* 2009]. Dans toutes ces études, la présence du contexte visuel aide à l'identification des sons d'environnement s'il est congruent. Notre but est d'évaluer si cette aide est suffisante pour contrebalancer les effets négatifs observés dans le cas où le signal sonore à identifier est bruité.

#### 5.2.5.4 Approche du contexte dans notre étude

Dans la section suivante, nous présentons une étude où nous avons choisi pour contexte une représentation visuelle simultanée de la source en elle-même. Ainsi, au contraire des travaux de [Ballas et Mullins 1991; Gygi et Shafiro 2011; Ozcan et van Egmond 2009] ou [Gérard 2004], le contexte tel que nous le définissons ne représente ni l'environnement ni un objet associé à la source par un lien sémantique. Enfin il n'est pas présenté en amorce comme dans [Schneider *et al.* 2008; Ballas et Mullins 1991] ou [Gérard 2004]. Notre considération du terme contexte est plus proche des travaux sur la redondance de cible et la facilitation intersensorielle (voir chap. 3, section 3.4.1.2, p. 49) tels qu'on peut les trouver par exemple dans [Suied *et al.* 2009].

## 5.3 Objectifs de l'étude et hypothèses

Nous avions pour objectif d'étudier trois aspects concernant l'identification des sons d'environnement :

a) l'effet de dégradation due au niveau de bruit environnant,

b) l'effet d'une représentation visuelle simultanée et de sa congruence sémantique,

c) la relation entre ces précédents facteurs.

Pour étudier les trois aspects précédemment cités, nous avons choisi d'utiliser des stimuli audiovisuels constitués d'une paire « son d'environnement + photographie représentant la source sonore ». Nous avions à disposition une collection de 140 sons d'environnement seuls préalablement étudiés par Bruno Giordano [Giordano *et al.* 2010]. Il nous a donc fallu dans un premier temps associer des photographies cohérentes pour correspondre au mieux à la source sonore de chaque son d'environnement de la collection. Ensuite, avant de pouvoir tester la dégradation sur l'identification des sons avec un faible RSB, nous avons d'abord vérifié que les images que nous avions sélectionnées étaient bien choisies, c'est-à-dire qu'elles étaient à la fois bien identifiées en elles-mêmes mais aussi qu'elles étaient bien associées au son qu'elles étaient sensées représenter, sans confusion entre les paires. L'étude est donc séparée en deux expériences distinctes, la première expérience servant à sélectionner l'ensemble de stimuli pour la seconde étude.

Nous supposions que, comme pour la perception de la parole, l'ajout d'un support visuel congruent serait une aide pour l'identification d'une source sonore, c'est-à-dire qu'il y aurait dans ce cas facilitation intersensorielle. Notamment nous voulions évaluer dans quelle mesure le contexte visuel pourrait compenser la perte d'informations auditives. De plus, en accord avec la théorie d'efficacité inverse, et d'après les études sur la perception audiovisuelle de la parole dans le bruit [Ross *et al.* 2007], nous pensions que cet effet de facilitation varierait proportionnellement à la perte de RSB (plus de bruit implique un plus grand effet du visuel).

## 5.4 Première expérience : sélection d'un ensemble de stimuli audiovisuels

Une première expérience a été menée afin de sélectionner un ensemble de stimuli audiovisuels facilement identifiables, constitués de sons d'environnement que les participants pourraient associer systématiquement à la même représentation visuelle de la source supposée de l'événement sonore. Les résultats de cette expérience ont été utilisés pour sélectionner des paires audiovisuelles comme stimuli pour l'expérience 2.

### 5.4.1 Sélection de sons d'environnement

Parmi les 140 sons de la collection de sons initiale, nous en avons pré-sélectionnés 50 de sorte que les stimuli aient tous un taux d'identification supérieur à 70 % (de 70 % à 100 %, moyenne 93 %, écart-type 9 %) d'après les résultats de [Giordano *et al.* 2010]. Les labels et les taux d'identification de ces stimuli sont reportés table 5.1. De plus, en vue de mesurer le temps de réaction nécessaire pour identifier les sons d'environnement en condition dégradée, il nous fallait une tâche où les participants n'auraient que deux choix possibles (*two-alternative forced*

*choice* ou *2AFC*). Nous avons choisi de demander aux participants de distinguer si la source sonore était produite par un être vivant, c'est-à-dire un animal ou un être humain, ou produite par un objet inanimé. Par abus de langage et pour simplifier la lecture du reste de l'étude, nous dirons d'un son qu'il est « *vivant* » s'il est produit par un être vivant et « *non vivant* » s'il est produit par un objet inanimé.

Plusieurs études ont déjà montré que ces deux catégories (« *vivant* » et « *non vivant* ») sont très distinctes, que ce soit en audio [Gérard 2004; Giordano *et al.* 2010] ou en vision [Laws 1999]. Ce genre de tâche est décliné sous différentes formes dans d'autres études sur les sons d'environnement. Par exemple Suied et al. [Suied *et al.* 2009] proposent une tâche de type *go/no-go* dans laquelle le participant doit appuyer sur le bouton *go* seulement si l'objet (sonore, visuel ou audio-graphique) ne présente aucun danger. Ce type de tâche a déjà été utilisé pour des expériences avec du visuel seul dans [Thorpe *et al.* 1996]. Dans le même ordre d'idée, Shneider et ses collègues proposent une tâche où le participant doit répondre si oui ou non l'objet représenté auditivement et/ou visuellement peut rentrer dans une boîte à chaussures [Schneider *et al.* 2008].

25 sons étaient donc produits par des sources « *vivantes* / animées » (22 vocalisations animales ou humaines et 3 sons produits par le corps d'animaux à savoir des sons de « grillons », d'une « personne se mouchant », ou d'une « mouche en vol ») tandis que les 25 autres sons étaient produits par des sources « *non vivantes* / inanimées » (à savoir des sons d'impact entre solides ou des sons de liquides). Nous avons évité les sons d'instruments de musique qui peuvent être catégorisés tantôt comme des sons d'environnement (« note isolée », « jouet pour enfant » ...) tantôt comme des sons exclusivement musicaux. Les sons de locomotion (« pas ») ont également été exclus de notre sélection car ils font intervenir en même temps un être humain et un solide inanimé selon le matériau du sol ( « gravier », « parquet », ...). Nous avons également retiré les bruits de bouches (« brossage de dents », « mastication ») jugés trop désagréables par les auditeurs. Enfin tous les sons devaient pouvoir être facilement associés à une représentation graphique sous forme de photographies. Nous avons donc supprimé les sons de la catégorie aérodynamique comme définie par [Gaver 1993] tels les sons de « souffle du vent ».

La durée des sons a été raccourcie à 3 s pour limiter les différences de durée d'un son à l'autre. La liste des 50 sons d'environnement utilisés est présentée en table 5.1.

**Table 5.1.** Ensemble des stimuli utilisés dans l'expérience 1. La colonne Label (*anglais*) correspond au label obtenu dans l'étude de [Giordano *et al.* 2010], tandis que la colonne Label (français) est une traduction française arbitraire pour faciliter la lecture de la partie résultat présentée en section 5.4.4. La colonne HNR(dB) indique le taux d'harmonicité maximal de chaque stimuli (calculé sur le son original sans bruit). La colonne %Correct Audio seul reprend les valeurs obtenues dans [Giordano *et al.* 2010]. Les colonnes %Correct Visuel seul et $p_{Assoc}$ Association A-V indiquent les résultats de l'expérience 1. La colonne Présence Expérience 2 indique quels sont les stimuli qui ont été conservés pour la seconde expérience de l'étude.

| **SONS VIVANTS** | | | | | | |
|---|---|---|---|---|---|---|
| Label (français) | Label (*anglais*) | HNR (dB) | %Correct Audio seul | %Correct Visuel seul | $p_{Assoc}$ Association A-V | Rôle dans l'Expérience 2 |
| Femme qui crie | *Screaming woman* | 34.3 | 95 | 90 | 100 | x |
| Mouche en vol | *Buzzing fly* | 12.6 | 95 | 100 | 100 | x |
| Chant du coq | *Crowing rooster* | 40.1 | 100 | 100 | 100 | x |
| Hennissement de cheval | *Neighing horse* | 12.5 | 100 | 100 | 100 | x |
| Grognement de cochon | *Grunting pig* | | 95 | 100 | 80 | Apprentissage |
| Cris de mouette | *Calling seagull* | 29.5 | 95 | 100 | 100 | x |
| Aboiement de phoque | *Barking seal* | 9.4 | 80 | 100 | 90 | x |
| Coassement de grenouille | *Croaking frog* | 10.5 | 100 | 100 | 90 | x |
| Caquettement de canard | *Quacking duck* | | 95 | 100 | 80 | Apprentissage |
| Rugissement de lion | *Roaring lion* | 9.0 | 100 | 100 | 100 | x |
| Femme haletante | *Gasping woman* | | 85 | 40 | 80 | NON |
| Gémissement de chien | *Whining dog* | 35.1 | 95 | 100 | 90 | x |
| Personne qui se mouche | *Blowing nose* | 12.6 | 100 | 100 | 100 | x |
| Bêlement de mouton | *Bleating sheep* | 15.5 | 100 | 100 | 90 | x |
| Chant de grillons | *Chirping cricket* | 22.4 | 85 | 100 | 100 | x |
| Croassement de corbeau | *Cawing crow* | 11.3 | 80 | 100 | 90 | x |
| Hurlement de loup | *Howling wolf* | 37.9 | 100 | 100 | 100 | x |
| Pleurs de bébé | *Crying baby* | 25.9 | 100 | 100 | 100 | x |
| Personne rotant | *Burping person* | | 100 | 0 | 95 | NON |
| Miaulement de chat | *Meowing cat* | 31.0 | 100 | 100 | 95 | x |
| Beuglement de vache | *Mooing cow* | 27.7 | 100 | 100 | 100 | x |
| Cri de l'aigle | *Calling eagle* | | 75 | 100 | 85 | Apprentissage |
| Barrissement d'éléphant | *Trumpeting elephant* | 19.4 | 100 | 100 | 95 | x |
| Homme qui tousse | *Coughing man* | 11.0 | 70 | 100 | 95 | x |
| Femme qui rit | *Laughing woman* | 21.2 | 90 | 100 | 100 | x |

suite page suivante...

... suite de la page précédente...

**SONS NON VIVANTS**

| Label (français) | Label (*english*) | HNR (dB) | %Correct Audio seul | %Correct Visuel seul | $p_{Assoc}$ Association A-V | Rôle dans l'Expérience 2 |
|---|---|---|---|---|---|---|
| Ballon que l'on gonfle | *Blowing balloon* | | 90 | 100 | 45 | NON |
| Eau qui bout | *Bubbling water* | 6.3 | 100 | 90 | 95 | x |
| Cloche | *Ringing bell* | 10.9 | 100 | 100 | 100 | x |
| Sonnette de vélo | *Ringing bike bell* | 14.3 | 90 | 100 | 100 | x |
| Bois que l'on scie | *Sawing wood* | 8.3 | 100 | 100 | 100 | x |
| Crépitement de feu de bois | *Crackling fire* | -3.9 | 75 | 100 | 85 | x |
| Trousseau de clé | *Jingling keys* | 6.1 | 100 | 100 | 95 | x |
| Dés qui roulent | *Rolling dice* | 3.8 | 75 | 100 | 100 | x |
| Klaxon de vélo | *Honking bike horn* | | 75 | 70 | 100 | Apprentissage |
| Eau versée dans un verre | *Pouring water* | 8.3 | 95 | 100 | 100 | x |
| Papier que l'on froisse | *Crumpling paper* | -2.9 | 95 | 100 | 95 | x |
| Swing de golf | *Swinging racket* | 7.3 | 85 | 100 | 100 | x |
| Sifflement de bouilloire | *Boiling kettle* | | 95 | 100 | 90 | NON |
| Goutte d'eau | *Dripping water* | 12.4 | 90 | 100 | 95 | x |
| Eau qui court (rivière) | *Running water* | -3.5 | 100 | 100 | 95 | x |
| Chasse d'eau | *Flushing toilet* | 8.0 | 100 | 100 | 100 | x |
| Epee affûtée | *Sharpening knife* | 12.1 | 90 | 90 | 100 | x |
| Touches de clavier | *Typing keyboard* | 2.0 | 100 | 100 | 100 | x |
| Monnaie qui tombe | *Dropping change* | -2.0 | 95 | 100 | 100 | x |
| Balle de ping-pong | *Bouncing ping pong ball* | 7.1 | 100 | 90 | 100 | x |
| Sifflet sans-gêne / cotillon | *Blowing party whistle* | 30.0 | 90 | 100 | 95 | x |
| Eclaboussement | *Splashing water* | | 90 | 60 | 100 | NON |
| Douche | *Flowing water* | -1.4 | 90 | 100 | 85 | x |
| Vagues caressant la plage | *Lapping water* | | 80 | 100 | 75 | Apprentissage |
| Coup de tonnerre | *Thundering thunder* | 1.9 | 100 | 100 | 100 | x |

### 5.4.2 Sélection de stimuli visuels associés

L'ensemble de stimuli visuels utilisé pour cette expérience comptait 60 images au total. 50 photographies ont été choisies pour être associées à l'ensemble de sons d'environnement précédemment décrit. Ces images proviennent de notre bibliothèque personnelle ou proviennent de ressources en ligne. Les images représentent directement la source sonore soit sur un fond blanc après détourage (20 images sur les 50), soit un arrière-plan texturé représentant une partie du contexte (exemple de l'herbe, la plage, pour 20 images) soit sur un arrière-plan coloré mais uni (exemple fond bleu : 10 images).

De plus, pour que les participants ne puissent pas procéder par élimination, nous avons ajouté 10 images supplémentaires, ne correspondant à aucune source sonore de la collection, et n'ayant aucun lien conceptuel avec les autres images. Afin de garder les mêmes proportions d'images que pour les 50 autres images, la moitié des images additionnelles représentait des animaux et l'autre moitié des objets inanimés, et quelques unes étaient sur fond blanc tandis que d'autres étaient sur fond coloré ou texturé. Toutes les images ont été redimensionnées de façon à avoir la même taille à l'affichage.

### 5.4.3 Evaluation pour valider le choix des stimuli

**Procédure** : L'expérience était divisée en deux étapes.

La première étape avait pour but de vérifier que les stimuli visuels, à savoir les photographies,

étaient correctement identifiés. Les 60 images étaient présentées simultanément sur une grille de 6 lignes par 10 colonnes, avec une taille réduite de $2.8 \times 2.1$ cm$^2$ (angle visuel d'environ $2° \times 1.5°$). L'ordre des images était tiré aléatoirement pour chaque participant de façon à mélanger les images d'agents animés et celles d'objets inanimés. Les participants avaient pour tâche de décrire librement ce qu'ils voyaient dans chaque image en écrivant en dessous un ou deux mots (verbes ou noms). La disposition des images pour cette étape est présentée en figure 5.4.

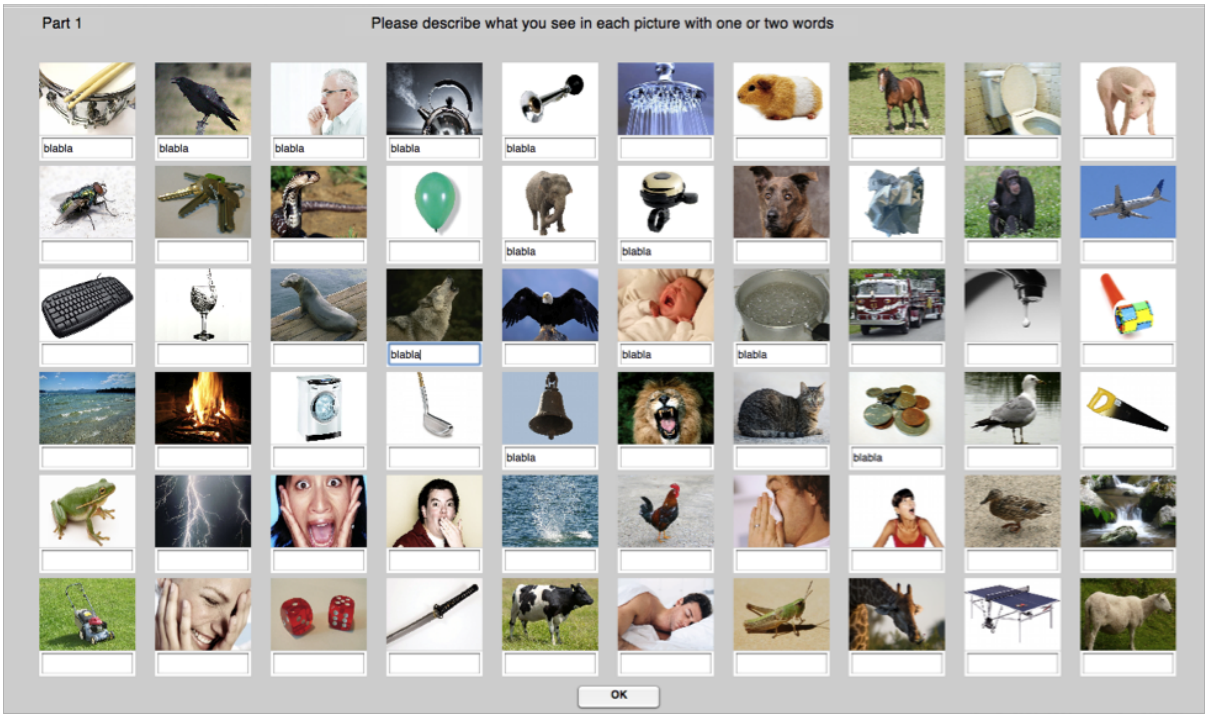



**Figure 5.4.** Présentation des photographies à décrire lors de l'étape 1.

Le but de la deuxième étape était de vérifier l'association entre les stimuli sonores et l'image choisie pour leur correspondre. A chaque essai, les participants entendaient un des 50 sons de l'ensemble de sons et devaient sélectionner, parmi les 60 images présentées, l'image qui correspondait au mieux au son entendu. Les participants validaient leur réponse en cliquant sur l'image. L'ordre des 60 images était le même que dans l'étape précédente de façon à ne pas perturber les participants mais aussi à limiter le temps d'expérience, chaque participant ayant déjà pris connaissances des images possibles au préalable. Chacun des 50 sons était présenté deux fois, dans un ordre aléatoire. Chaque participant répondait donc à 100 essais. Il ne recevait aucune information sur la justesse de ses réponses durant l'expérience. Au total, l'expérience totale (étapes 1 et 2) durait approximativement 20 minutes par participant.

**Participants** : Dix personnes (3 femmes, 7 hommes) ont participé individuellement à cette expérience. Toutes avaient une vue correcte ou corrigée et aucune n'a rapporté avoir de problème auditif.

### 5.4.4 Résultats et discussion

Cette expérience avait pour but de sélectionner un ensemble de stimuli audio et visuels appariés de sorte que les stimuli soient bien reconnus dans chaque modalité et que les éventuelles erreurs d'identification ne puissent être dues ni à un problème d'association ni à un problème de cohérence entre l'élément sonore et l'élément visuel de chaque paire. La première étape permettait de mesurer le taux d'identification des stimuli visuels (%Correct Visuel seul). La seconde étape mesurait le taux d'association correcte entre un stimulus sonore et l'image choisie pour le représenter ($p_{Assoc}$). Ces données sont disponibles dans le tableau 5.1.

**Calcul du taux d'identification des images seules %Correct Visuel seul**

Nous avons considéré comme correctes les réponses qui correspondaient au label obtenu pour les sons d'environnement d'après [Giordano *et al.* 2010]. En reprenant les règles pour le nommage des sons d'environnement décrites dans [Marcell *et al.* 2000], nous avons accepté des termes sus-spécifiques (comme « cobra » à la place de « serpent ») s'ils correspondaient bien à l'élément principal de l'image. Les termes sous-spécifiques, ne décrivant pas suffisamment l'image, ont été comptés comme faux (exemple : « homme » à la place de « homme endormi »).

Seules deux images ont été mal interprétées avec un taux d'identification inférieur à 75 % : une « personne rotant » ("*burping person*", 0 % d'identification) que les participants ont pris pour une « personne baillant » à 30 % ou pour un « homme surpris » à 70 % (voir figure 5.5a) ; et l'image d'une « femme haletante » ( "*woman gasping*") reconnue à seulement 40 % (sous les termes « surprise » et « admiration » qui ont été acceptés) confondue avec la « peur » à 20 % ou identifiée sous le label imprécis de « femme » (40 %) (voir figure 5.5b).

**(a)** « personne rotant ».

**(b)** « femme haletante ».

**Figure 5.5.** Deux images mal interprétées par les participants donc retirées du set.

**Taux d'association son-images**

La seconde étape de l'expérience permettait de mesurer le taux d'association son-image ($p_{Assoc}$) défini comme le nombre de fois qu'un son était associé avec le stimulus visuel que nous avions choisi, arbitrairement, comme représentation visuelle de la source sonore. Parmi les stimuli, une source sonore sensée être « *non vivante* » (« sifflement de bouilloire », exemple audio 5.1) a été associée une fois avec l'image d'un animal (« mouche en vol », figure 5.6). Pour éviter toutes confusions futures entre les deux catégories, nous avons retiré les deux paires son+image relatives à la « mouche » et à la « bouilloire ». Nous avons également observé

5.1

plusieurs confusions entre les sources sonores relatives à des sons de liquide (« eau versée dans un verre » audio 5.2, « vagues caressant la plage » audio 5.3, « rivière » audio 5.4, « douche » audio 5.5). Pour cette raison nous avons retiré une des paires son+image de source liquide, en choisissant celle avec le taux d'association $p_{Assoc}$ le plus faible (« vagues caressant la plage » ou "*lapping water*" , $p_{Assoc} = 75$ %). Enfin l'image d'un « ballon que l'on gonfle » n'a jamais été correctement associé au son correspondant (exemple audio 5.6) car les participants associaient systématiquement ce son à l'image d'un « avion », image qui faisait pourtant partie des images additionnelles sans son associé (figure 5.7).

5.2 5.3 5.4 5.5 5.6

**Figure 5.6.** Image d'une mouche, associée par erreur au son de « sifflement de bouilloire ».

**Figure 5.7.** Image d'un avion, associée par erreur au son de « ballon que l'on gonfle ».

Au final de cette étape de validation, sept paires de stimuli audiovisuels ont été exclues dû à des mauvaises identifications de la représentation visuelle (2 paires) ou de la mauvaise association entre la composante visuelle et sonore de la paire (5 paires). Pour garder le même nombre de paires audiovisuelles dans chaque catégorie (agents animés vs. objets inanimées), nous avons également retiré 3 paires de la catégorie « *vivant* » supplémentaires, en prenant celles pour lesquels le $p_{Assoc}$ était le plus faible : « cochon qui grouine » ("*grunting pig*", $p_{Assoc} =$ 80 %), « canard caquetant » ("*quacking duck*", $p_{Assoc} = 80$ %) et « aigle qui glatit » ("*calling eagle*", $p_{Assoc} = 85$ %). Au final nous avons gardé 40 paires son+image représentant des sources différentes (20 stimuli « *vivants* » et 20 « *non vivants* ») avec un taux d'identification des images seules allant de 90 % à 100 % (moyenne 99 %) et un taux d'association entre l'image et le son allant de 85 % à 100 % ( moyenne 93.4 %).

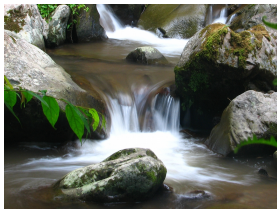
## 5.5 Deuxième expérience : identification des sons d'environnement avec un rapport signal-sur-bruit dégradé

Cette deuxième expérience avait pour but de mesurer l'effet du contexte visuel sur l'identification de sons d'environnement présentés en présence de bruit masquant de niveau plus ou moins élevé. Cette condition dégradée peut être vue comme la simulation d'un environnement acoustique chargé perturbant l'écoute de certains sons spécifiques. Comme expliqué précédemment, nous avons basé cette étude sur une tâche « *vivant / non vivant* » où les participants devaient identifier chaque son pour le classer dans l'une de ces deux catégories selon l'origine du son. Trois

conditions de combinaisons audiovisuelles (AV) ont été testées : audio + représentation visuelle congruente de la source sonore (**AVc**), audio + représentation visuelle incongruente avec une image issue de l'autre catégorie (**AVi**), audio + représentation d'une image neutre abstraite ne représentant aucune source sonore (**AVn**). Nous avons également testé une condition contrôle avec de l'audio seul (**A**). Ainsi par exemple, si le son à identifier est celui d'une « femme qui crie » (exemple audio 5.7), il est classé dans la catégorie « *vivante* ». L'image pour la condition **AVc** est celle d'une femme qui crie (figure 5.8a) tandis qu'une image possible pour la condition **AVi** est tirée aléatoirement parmi les images des sources non vivantes (par exemple une rivière qui coule, figure 5.8b). Une image neutre pourrait alors être celle présentée figure 5.8c.

5.7

**(a)** Image congruente.



**(b)** Image incongruente.

**(c)** Image abstraite.

**(d)** Audio seul.

**Figure 5.8.** Exemple des associations visuelles possibles pour la source sonore vivante « femme qui crie ».

Nous partons de l'hypothèse que la modalité visuelle influence l'identification des sources sonores, en condition d'écoute normale mais aussi dégradée. De plus, cette influence visuelle serait d'autant plus importante que le RSB serait faible, limitant alors l'identification des sources sonores pour une présentation uniquement auditive. En effet, plusieurs études ont observé que l'ajout de la modalité visuelle renforce l'intelligibilité de la parole dans le bruit, et ce d'autant plus que le niveau de bruit est élevé [Sumby et Pollack 1954]. Cela est cohérent avec les théories récentes sur l'intégration multisensorielle qui suggèrent que plus une modalité donne des informations ambigües ou inexploitables, moins l'on se réfère à elle [Ernst et Bülthoff 2004; Stein et Meredith 1993]. Dans notre cas, plus les sons sont dégradés et difficiles à identifier, plus notre interprétation basée sur l'écoute est limitée, et donc, nous sommes plus amenés à nous appuyer sur la modalité visuelle. Pour vérifier cela nous avons utilisé 8 niveaux de bruit différents. Le protocole expérimental détaillé ci-après suit donc un plan factoriel croisé avec 8 RSB $\times$ 4 conditions AV $\times$ 2 catégories (« *vivant* / *non vivant* »).

### 5.5.1 Sélection des stimuli sonores et visuels

Bien que validés par l'expérience précédente, les stimuli de cette deuxième expérience sont légèrement différents. D'une part, pour l'audio, nous avons ajouté des versions bruitées des sons utilisés pour la première expérience. D'autre part, pour le visuel, nous avons ajouté des images, abstraites, pour une condition d'association neutre entre le contexte visuel et les sons à identifier.

Nous avons utilisé les 40 sons d'environnement sélectionnés à partir des résultats de l'expérience précédente. Ces sons (20 vivants, 20 non vivants) étaient tous identifiables seuls [Giordano

*et al.* 2010] et étaient associés sans ambiguïté à l'image correspondante (voir expérience 1). Tous les stimuli ont d'abord été égalisés en niveau global RMS avant d'être traités pour obtenir différentes conditions de RSB. Ainsi 7 autres versions de chaque son ont été obtenues en ajoutant, sur la durée totale du stimulus (3 s), un bruit blanc de niveau tel que le RSB variait de −18 dB à 0 dB par pas de 3 dB.

Comme les niveaux ont été égalisés en niveau global (RMS), nous avons regardé s'il y avait une différence entre les deux catégories sur le niveau maximum ponctuel (pic-à-pic). Nous avons relevé une différence de niveau maximal entre les 2 catégories : le pic de niveau des sons « *non vivants* » étaient en moyenne plus élevé que le niveau de crête des sons « *vivants* » d'environ 8 dB en moyenne (t(38), $p$ <0.001, *t-test*). Nous avons alors calculé la position du centre temporel de chaque stimuli que nous avons défini comme le barycentre de chaque vecteur signal. Il est plus court pour les sons « *non vivants* » (en moyenne 1.28 s) que pour les sons « *vivants* » (moyenne = 1.54 s) (t(38) = 2.98, $p$ <0.01). On peut donc s'attendre à ce que les sons d'environnement « *non vivants* » passent plus rapidement au dessus du niveau de bruit de fond. Si c'est le cas, les indices acoustiques pertinents pour identifier la source sonore seront disponibles plus vite pour les sons de sources « *non vivantes* ». On peut donc supposer que ces sons seront identifiés plus rapidement.

40 stimuli visuels (20 vivants, 20 non vivants), correspondant à chacun des stimuli audio ont été sélectionnés d'après les résultats de l'expérience précédente. Nous avons ajouté à cet ensemble de stimuli 24 images abstraites pour une condition audio + visuel neutre. Comme dans l'expérience de [Schneider *et al.* 2008], ces images étaient des images synthétiques avec différentes traces colorées et ne pouvaient pas être clairement associées à aucune des sources sonores. Des exemples de stimuli visuels employés dans cette étude sont présentés en figure 5.9.

**(a)** Image vivante. **(b)** Image non vivante. **(c)** Image abstraite.

**Figure 5.9.** Exemples d'images de chaque catégories de stimuli visuels.

### 5.5.2 Procédure expérimentale

La tâche était une tâche de catégorisation sémantique similaire à celle utilisée dans les travaux de Gérard [Gérard 2004, p. 170] : les participants devaient décider aussi vite que possible si un son était de type « *vivant* » produit par un animal ou un être vivant ou un son de type « *non vivant* » produit par un objet inanimé. Les participants répondaient en pressant une des deux touches du clavier sur lesquelles étaient collées des étiquettes « liv. » (pour *living*) et « non liv. » (pour *nonliving*) [1]. Les deux touches choisies étaient les touches 'n' et 'v' et la

1. L'expérience était menée sur des participants anglophones.

correspondance entre ces touches et les deux catégories était contrebalancée entre les participants (touche 'n' et 'v' pour la moitié, touche 'v' et 'n' pour l'autre moitié). Ils étaient également prévenus que les sons pouvaient être, éventuellement, masqués par du bruit et que les images présentées en parallèle des sons pouvaient ou non représenter la source produisant le son. Il leur était explicitement demandé de regarder l'écran attentivement pendant l'écoute, mais de ne déterminer la catégorie qu'à partir du son. Dès que le participant pressait une des deux touches de réponses, la lecture du son était arrêtée et l'image retirée. Un nouvel essai débutait automatiquement après une pause de 700 ms (figure 5.10). Contrairement aux études de [Gérard 2004], les participants ne recevaient aucun retour sur l'exactitude de leur réponse.

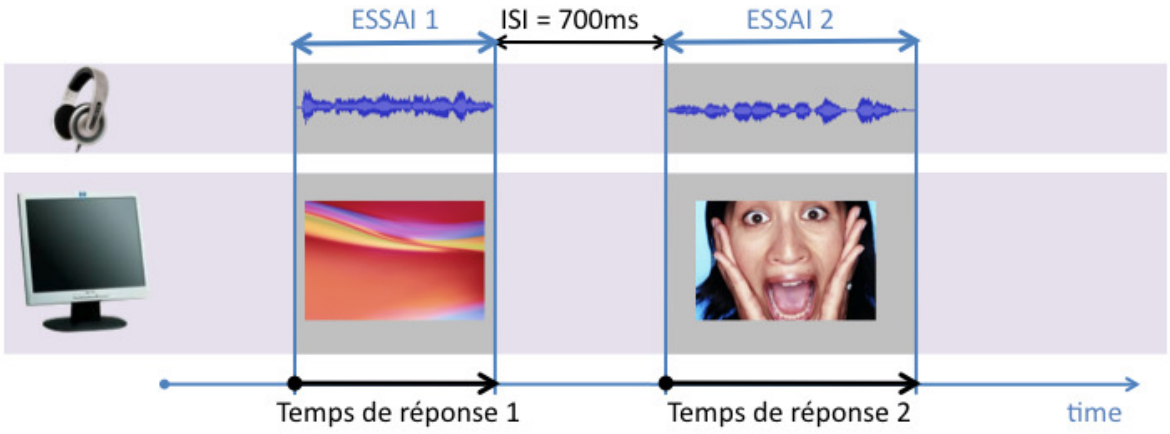



**Figure 5.10.** Scénario de l'expérience 2. Ici l'essai 1 est un exemple de la condition AV neutre.

**Variables et plan d'expérience** : Pour cette expérience principale, nous avons utilisé un plan d'expérience de type factoriel entre les 8 *conditions de RSB* et les 4 *combinaisons de modalité audiovisuelles*. Pour chacune de ses 8×4 conditions, 12 sources de chaque catégorie étaient choisies aléatoirement parmi les 20 sources possibles. Nous voulions en effet limiter le nombre de répétitions d'un même stimulus durant l'expérience pour éviter un effet de familiarisation et également limiter le nombre d'essais expérimentaux pour chaque participants. Au final, chaque participant réalisait la tâche sur 8 conditions RSB × 4 conditions AV × 2 catégories × 12 sources = 788 essais répartis aléatoirement en 16 sessions de 48 essais. Les participants pouvaient prendre des pause entre chaque session. Avant l'expérience principale, les participants se familiarisaient avec la tâche sur 16 essais d'entraînements. Pour cette étape d'entraînement, nous avons employé 6 paires son+image non sélectionnées pour l'expérience principale. L'expérience durait au total environ 45 min.

**Participants** : 19 personnes (âge moyen 24 ans, 9 hommes, 17 droitiers) ont participé à cette expérience. Aucune d'entre elles n'avait participé à l'expérience 1. Toutes avaient une vue normale ou corrigée sans historique de problème de vision. Les participants passaient auparavant un test audiométrique attestant qu'ils n'avaient pas de problème auditif (niveau audiométrique normal (< 20 dB HL) par bande d'octaves entre 250 et 8000 Hz).

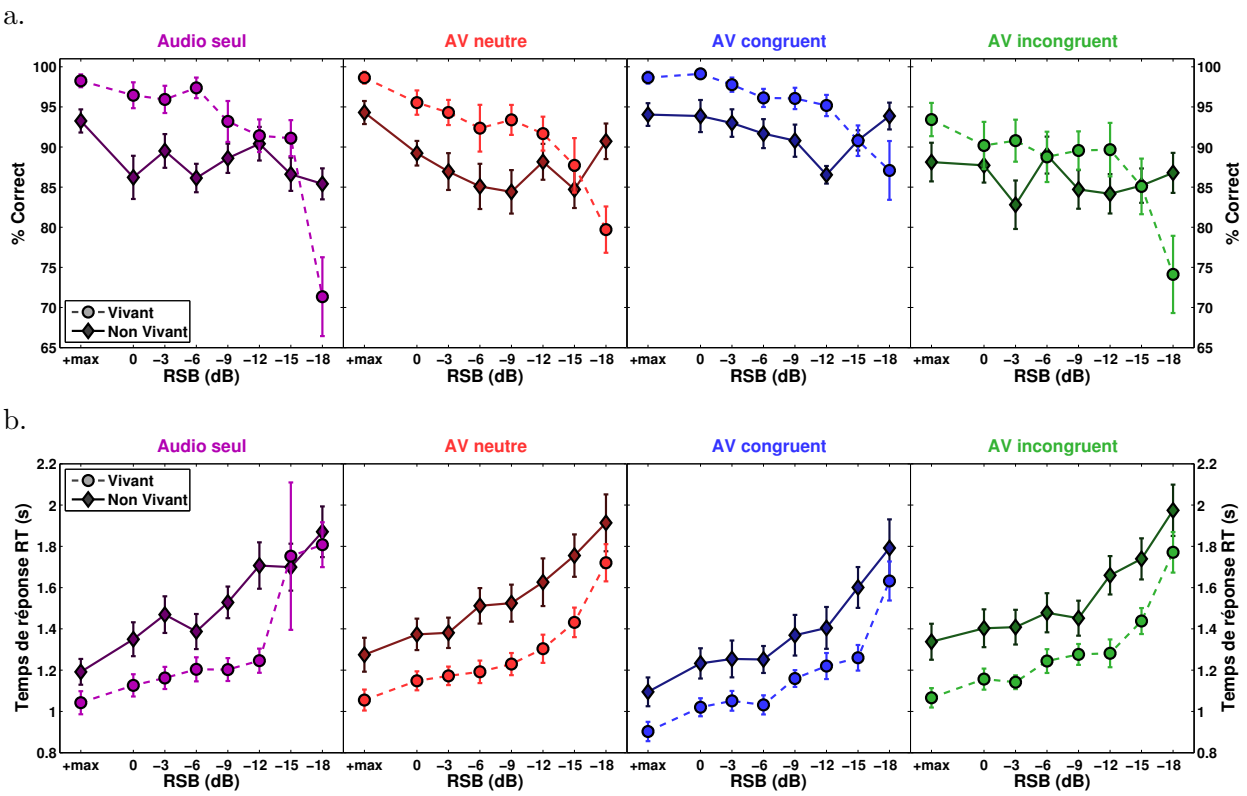



**Figure 5.11.** Résultats des différentes conditions expérimentales en fonction du rapport signal-sur-bruit (RSB). +max= sans bruit. Les barres d'erreur représentent l'erreur standard moyenne. a) Taux de réponses correctes. b) Temps de réponse moyen sur les réponses correctes seulement.

### 5.5.3 Résultats

La figure 5.11 présente les résultats moyens sur l'ensemble des participants de deux paramètres mesurés : le taux de réponses corrects (%Correct) et les temps de réponses (RT). Nous observons que plus le RSB diminue, plus le %Correct diminue, et plus les RT s'allongent, et ce pour toutes les catégories de son, et toutes les conditions de congruence audiovisuelles. On peut également observer une nette différence entre les conditions de congruence : la condition **AVc** conduit à des RT plus courts que ceux de **A** ou **AVi**.

Enfin, on remarque une différence prononcée dans les résultats obtenus pour les sons « *vivants* » et les sons « *non vivants* ». Cette dernière observation ne répond pas directement à nos questions de recherche concernant l'effet du bruit et du contexte visuel. Toutefois des analyses plus poussées ont été menées concernant cette différence « *vivants / non vivants* » pour apporter des connaissances supplémentaires sur notre perception des sons d'environnement en général (section 5.5.4).

**Table 5.2.** Détail des résultats de %Correct moyennés sur les différentes conditions de RSB (entre parenthèses l'écart-type).

| | | Congruence audiovisuelle | | | |
|---|---|---|---|---|---|
| | | **A** | **AVn** | **AVc** | **AVi** |
| Taux de réponses correctes (*en*%) | Vivant | 91.9 (1.7) | 91.7 (1.6) | 95.1 (1.0) | 87.7 (2.5) |
| | Non vivant | 88.3 (1.0) | 87.9 (1.1) | 91.8 (0.9) | 86.1 (1.7) |
| | Global | 90.1 (1.2) | 89.8 (1.2) | 93.5 (0.8) | 86.9 (2.0) |

#### 5.5.3.1 Taux d'identification

Le détail des résultats de %Correct peut être trouvé en table 5.2. Une analyse de variances à mesures répétées (RM-ANOVA) a été effectuée, avec comme paramètres intra-participants : la congruence audiovisuelle (**A**, **AVn**, **AVc**, **AVi**), le RSB (8 niveaux de bruits) et la catégorie (« *vivants, non vivants* »). L'analyse a révélé un effet significatif du RSB ($F(7,126)=26.76$, $p \ll 0.001$), un effet significatif de la congruence audiovisuelle ($F(3,54)=13.48$, $p \ll 0.001$) et un effet significatif de la catégorie ($F(1,18)=6.36$, $p = 0.021$). De plus, les interactions RSB × congruence ($F(2,378)=2.64$, $p \ll 0.001$) et RSB × catégorie ($F(7,126)=11.59$, $p \ll 0.001$) sont significatives. Ainsi l'influence d'un même niveau de bruit sur l'identification des sons dépend de la catégorie de son et du contexte visuel. Les autres interactions, congruence × catégorie et RSB × congruence × catégorie, ne sont pas significatives.

Ainsi, on observe :

**Un effet croissant du niveau de bruit** : le taux de réponses correctes %Correct diminue avec le RSB. Cette baisse d'identification est plus importante pour les sons « *vivants* » et elle a lieu même en présence de visuel congruent ce qui implique que le visuel ne suffit pas à compenser totalement la perte auditive ;

**Un impact de la congruence sémantique entre le visuel et l'audio** : l'identification est améliorée en condition audiovisuelle congruente **AVc** (seulement 6.54 % d'erreur) par rapport à toutes les autres conditions ($p \ll 0.001$), et en particulier comparé à une condition avec seulement l'audio **A** (9.93 % d'erreur) ou un rendu visuel sans relation **AVn** (10.20 % d'erreur). En revanche, dans le cas d'un conflit sémantique entre l'audio et le visuel, comme dans le cas **AVi** où le visuel indiquait la mauvaise catégorie, l'identification de la source sonore diminue (13.12 % d'erreurs) ;

**Une asymétrie entre les sons « *vivants* » et « *non vivants* »** : en moyenne les participants ont fait moins d'erreur avec les sons « *vivants* »(8.41 %) qu'avec les sons « *non vivants* »(11.45 %). Toutefois cette asymétrie ne s'observe que pour des niveaux de bruits faibles et moyens (RSB $>-15$ dB), l'asymétrie entre les deux catégories est inversée pour les RSB égaux à $-15$ dB et à $-18$ dB.

**Table 5.3.** Détail des résultats de RT moyennés sur les différentes conditions de RSB. Entre parenthèse figure l'écart-type.

| | | Congruence audiovisuelle | | | |
|---|---|---|---|---|---|
| | | **A** | **AVn** | **AVc** | **AVi** |
| RT (en ms) | Vivant | 1318 (81) | 1282 (51) | 1160 (48) | 1297 (52) |
| | Non vivant | 1525 (80) | 1545 (85) | 1375 (81) | 1557 (85) |
| | Global | 1422 (78) | 1413 (66) | 1267 (63) | 1427 (66) |

#### 5.5.3.2 Temps de réaction

Les temps de réponse n'ont été analysés que pour les réponses correctes. Le détail des RT pour les différentes conditions de congruence et de catégorie, et pour tout niveau de bruit est présenté table 5.3. Une RM-ANOVA reprenant les mêmes facteurs que pour l'analyse des %Correct a été utilisée pour l'analyse des RT. L'analyse a révélé des effets significatifs du RSB ($F(7,126)= 51.60$, $p \ll 0.001$), de la congruence audiovisuelle ($F(3,54)= 30.48$, $p \ll 0.001$) et de la catégorie ($F(1,18)=41.10$, $p \ll 0.001$). En revanche, l'analyse n'a révélé aucune interaction significative entre les facteurs ($F's \leq 1.46$, $p's \geq 0.09$). Ainsi, par exemple, le gain en temps pour la condition congruente ne dépend pas du niveau de bruit et le rôle du visuel n'augmente pas avec la baisse de fiabilité auditive due au bruit.

Comme pour l'analyse des %Correct, on observe sur les RT :

**Un effet croissant du niveau de bruit** : quelque soit la catégorie et la condition de congruence, plus le RSB diminue et plus les RT sont longs. Il faut donc plus de temps pour reconnaître une source sonore dans un environnement bruyant ;

**Un impact de la congruence sémantique entre le visuel et l'audio** : les RT les plus courts sont obtenus pour une présentation audiovisuelle congruente **AVc**. En revanche, les RT ne sont pas plus longs dans le cas incongruent que dans le cas audio seul : d'un point de vue temps de réponse, tout se passe comme si l'image abstraite ou incongruente n'était pas traitée ;

**Une asymétrie entre les sons « *vivants* » et « *non vivants* »** : dans toutes les conditions de congruence et de RSB, les participants ont identifié les sons « *vivants* » plus rapidement que les sons « *non vivants* » (avec des moyennes respectives de 1264 ms et 1500 ms).

### 5.5.4 Analyse supplémentaire concernant l'asymétrie « vivant / non vivant »

Les performances des participants révèlent une forte différence entre la perception des sons « *vivants* » et celle des sons « *non vivants* ». Relevée par de nombreuses études antérieures [Lewicki 2002], la distinction entre ces deux catégories est intriguante.

Nous voulions évaluer dans quelle mesure cette distinction pouvait s'expliquer uniquement par l'information acoustique. Nous avons donc effectué une analyse acoustique pour estimer

l'influence du RSB et du contexte visuel sur le traitement perceptif des informations acoustiques utilisées dans la distinction de catégorie.

En particulier, nous nous sommes concentrée sur la notion d'***harmonicité***. En effet, ce paramètre nous semblait être la propriété acoustique qui séparaît le mieux les sons « *vivants* » des sons « non vivants », dans le sens où les sons « *vivants* » sont constitués en grande majorité de vocalisations dont le contenu est harmonique. Nous avons donc comparé l'harmonicité de chaque stimulus sonore avec la probabilité que les participants l'identifient comme « *vivant* » ($p_{vivant}$). Notre hypothèse est que plus le signal est harmonique et plus il ressemble à une vocalisation et, donc plus il est associé à une source « *vivante* ».

#### 5.5.4.1 Calcul du rapport d'harmonicité (HNR)

Nous avons utilisé le logiciel Praat [2] décrit dans [Boersma 2001] pour calculer le HNR de chaque stimulus sonore avant l'ajout de bruit. Le calcul a été effectué avec les valeurs par défaut [3] de la fonction `To Harmonicity`. L'harmonicité d'un son variant au cours du temps, nous avons gardé d'une part le HNR moyen et d'autre part la valeur maximale du HNR de chaque stimulus. Les analyses reportées ci-après concernent uniquement la valeur maximale du HNR, toutefois des observations équivalentes sont obtenues pour la valeur moyenne. La valeur maximale des HNR est rapportée en table 5.1, page 101.

On remarque que la moyenne des HNR des sons « *vivants* » ($\sim$ 21.45 dB) est plus grande que celle des sons « *non vivants* » ($\sim$ 6.24 dB). Un test de Student, avec données non appariées, confirme que cette différence est significative ($t(38)=5.18$, $p \ll 0.001$). Cela tend à confirmer que le HNR serait un bon indice acoustique sur lequel se baser pour effectuer la distinction entre les catégories « *vivante* » et « *non vivante* ». Pour vérifier dans quelle mesure les participants ont pu se fier à ce paramètre, nous avons calculé pour chaque son, « *vivant* » ou « *non vivant* », la probabilité que les participants l'identifient comme « *vivant* » ($p_{vivant}$). Le $p_{vivant}$ de chaque son est calculé comme le nombre total de réponses « *vivant* » obtenues pour ce son divisé par le nombre total d'apparitions de ce son. Par exemple, si un son est catégorisé 20 fois sur 100 essais comme « *vivant* » (et 80 fois sur 100 comme « *non vivant* »), son $p_{vivant}$ est de 0.20. Nous avons ensuite cherché comment évolue le $p_{vivant}$ en fonction du HNR.

Pour une condition sans bruit et sans contexte visuel, nous obtenons des probabilités de réponses décrites à la figure 5.12. On observe que les sons « *vivants* » sont correctements reconnus à plus de 80 %, c'est-à-dire que le $p_{vivant}$ de ces sons est supérieur à 0.8. De plus, 93.4 % des sons dont le HNR est supérieur à 14 dB sont « *vivants* » et reconnus comme tels par les participants, l'exception étant le son de « sifflet » (« *non vivant* ») dont le HNR est particulièrement élevé (30.0 dB) mais qui est reconnu à 100 % comme « *non vivant* » dans cette condition. Enfin, tous les sons dont le HNR est inférieur à 9 dB sont des sons « *non vivants* » et sont reconnus comme tels, à l'exception de deux sons (« eau ruisselante » et « bois qu'on coupe ») qui sont identifiés le plus souvent comme « *vivants* ». Au final, 12 sons « *vivants* », reconnus comme tels, ont un HNR supérieur à 14 dB et 13 sons « *non vivants* », reconnus comme tels, ont un HNR inférieur

2. Praat est librement distribué à l'adresse : http://www.fon.hum.uva.nl/praat/

3. Ces valeurs par défaut correspondent à : une fenêtre d'analyse (*Minimum Pitch*) de 75 Hz, un intervalle (*Time step*) de 0.01 s, un seuil de silence (*Silence Threshold*) de 0.1, et un nombre de périodes par fenêtre (Number of periods per wondow) de 4.5.

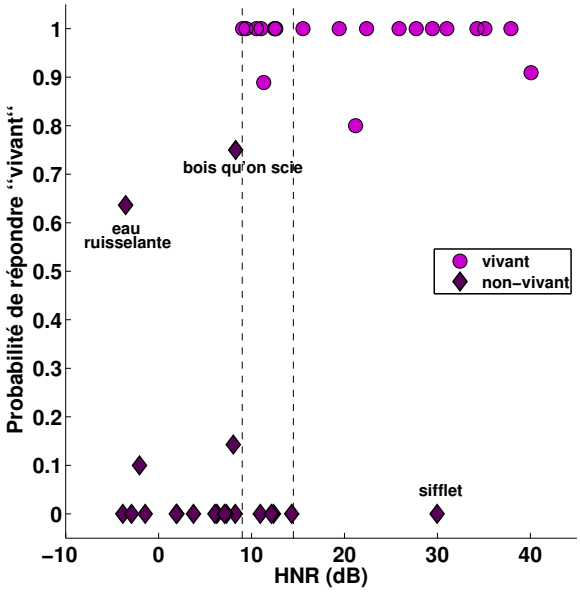


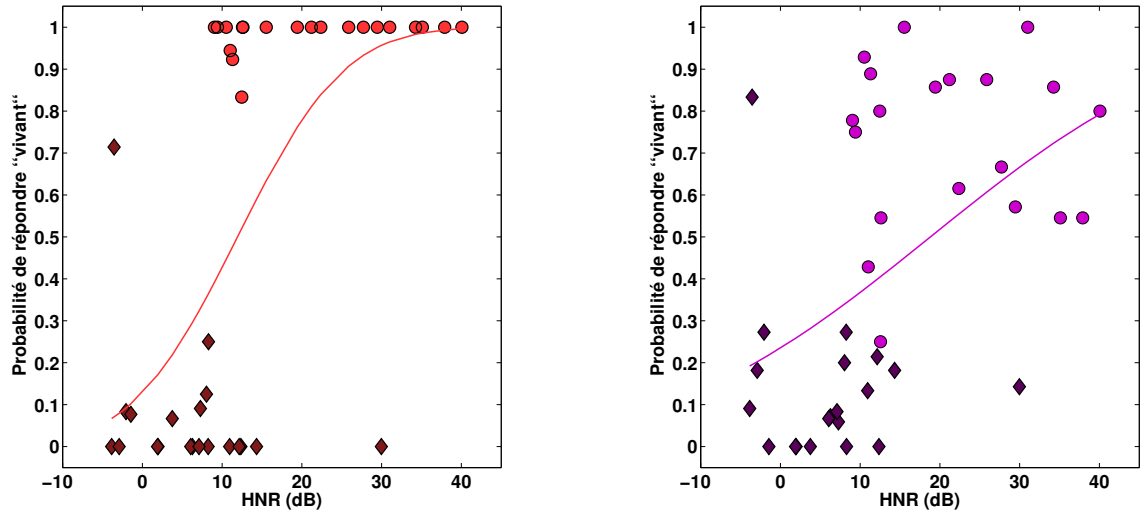

**Figure 5.12.** Mise en relation entre le choix des participants et le rapport d'harmonicité des stimuli pour une condition sans bruit ajouté (+max dB) et sans contexte visuel (Audio seul).

à 9 dB. Ces 25 sons, soit 60 % des stimuli, ont pu être catégorisés en ne tenant compte que du HNR.

#### 5.5.4.2 Effet du bruit et du contexte visuel sur la prise en compte du HNR

Nous venons de voir que les participants se sont très probablement appuyés sur l'indice d'harmonicité HNR pour réaliser la tâche de catégorisation « *vivant / non vivant* ». Nous cherchons maintenant à savoir si le contexte visuel et le bruit ajouté ont modifié la façon dont les participants ont pu s'appuyer sur cet indice acoustique.

Pour ce faire, nous avons calculé un modèle de régression généralisé de type *probit* (pour *probability unit*) pour chacunes des 32 conditions (8 niveaux de bruit × 4 congruences audio-visuelles). Ce modèle, calculé à partir des 40 observations (HNR, $p_{vivant}$), permet de prédire la probabilité de répondre « *vivant* » connaissant le HNR d'un son. Quelle que soit la condition de niveau de bruit, le modèle probit était toujours calculé à partir des HNR des sons originaux avant ajout de bruit. Cela revient à supposer que les participants étaient toujours capables de séparer le son du bruit blanc ajouté, et qu'ainsi ils étaient capables d'estimer l'harmonicité du son entendu indépendamment du niveau de bruit de fond. Cette hypothèse nous permet d'évaluer dans quelle mesure les participants se sont fiés au paramètre d'harmonicité pour faire leurs choix. Le modèle probit n'indique pas, lorsque les participants se fient moins à l'harmonicité, si cela est dû à une difficulté préalable de séparer le son à identifier du bruit ajouté, ou dû à d'autres facteurs. Il permet finalement d'obtenir une fonction qui approxime au mieux les points (HNR, $p_{vivant}$) tels que présentés sur un graphe similaire à la figure 5.12. Deux exemples de modèle probit sont présentés figure 5.13.




**(a) AV neutre** ; Sans bruit ajouté.

**(b) A seul** ; RSB = -18 dB.

**Figure 5.13.** Deux modèles probit et l'ensemble des 40 observations associées. Plus la pente est forte et plus il est probable que les participants se soient fiés au HNR pour distinguer les deux catégories de sons (vivant ∘, non vivant ⋄).

Lorsque les $p_{vivant}$ sont nuls ou réciproquement valent 1, cela signifie que les participants identifient sans équivoque les sons comme « *non vivants* », réciproquement « *vivants* », et donc qu'ils ont eu suffisamment d'indices acoustiques pour les identifier. La pente du modèle probit sera alors forte, indiquant à quel point la probabilité de répondre « *vivant* » change rapidement en fonction du HNR (figure 5.13a). Au contraire, un point situé au centre verticalement ($p_{vivant} \simeq 0.5$) indique que les participants répondent au hasard pour catégoriser ce son (2 réponses possibles induisent un niveau de chances égal à 50 %) : ils n'ont pas pu extraire d'indice acoustique suffisamment pertinent et en particulier n'ont pas pu se fier au HNR. Or plus les points sont au centre et plus la courbe obtenue par le modèle probit est plate (avec une pente faible) (figure 5.13b). La pente du modèle probit peut donc être considérée comme une mesure de la sensibilité au paramètre d'HNR.

Nous avons calculé la pente des 32 modèles obtenus (un modèle par condition) avec comme variable indépendante les HNR et comme variable dépendante les valeur de $p_{vivant}$. La figure 5.14 présente la pente des différents modèles en fonction du RSB et de la congruence AV. Une ANOVA à deux facteurs (RSB et congruence) a été réalisée pour analyser les différences de pente des modèles probit dans les différentes conditions. Quelle que soit la congruence AV, une diminution du RSB implique une diminution de la pente, indiquant une baisse significative de la sensibilité au HNR ($F(7,21)=5.84$, $p \ll 0.001$). Le paramètre de congruence AV a aussi un effet significatif sur cette sensibilité ($F(3,21)=3.62$, $p = 0.03$). En particulier, la pente est plus prononcée pour la condition **AVc** (en moyenne 0.113) que pour les autres conditions. La condition $AVc$ étant la condition avec les pentes les plus basses. Un test post hoc par paire a montré que la seule différence significative entre les conditions AV était en fait entre les conditions **AVc** et **AVi**

(t(7)=5.94, $p \ll 0.001$). Ainsi les résultats indiquent que les participants se sont plus fiés au HNR lorsqu'une présentation visuelle congruente était ajoutée que lorsque la présentation visuelle indiquait l'autre catégorie. Aucune interaction significative entre les paramètres de RSB et de congruence AV n'est observable.

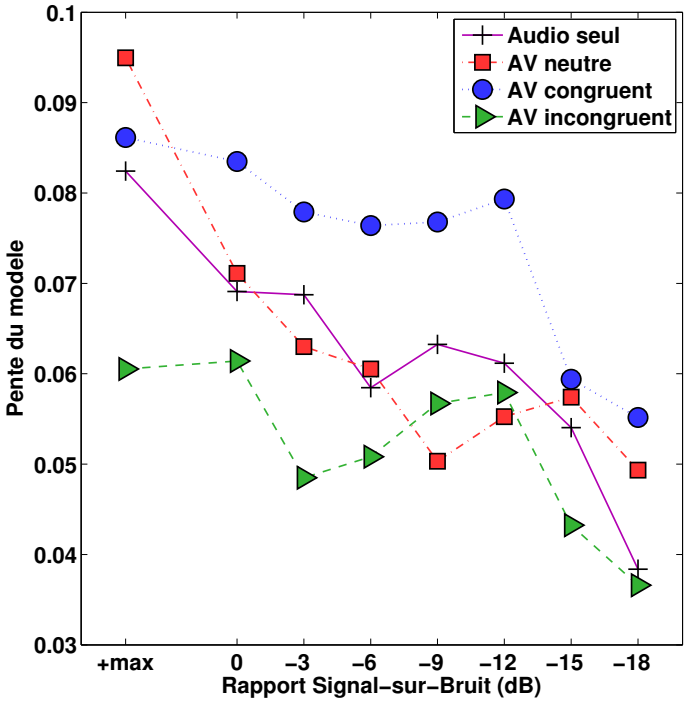



**Figure 5.14.** Pente des modèles probit pour estimer la probabilité de répondre « *vivant* » dans les différentes conditions expérimentales. Les pentes fortes indiquent une sensibilité plus grande au rapport d'harmonicité HNR du signal sonore.

#### 5.5.4.3 Etude d'un autre facteur acoustique : mesure de la centroïde spectrale

Comme montré dans les études de [Misdariis *et al.* 2010], la centroïde spectrale (moyenne pondérée des fréquences présentes dans le signal à un instant donné) est également un facteur prépondérant dans l'identification des sons d'environnement. Nous avons voulu vérifier si ce facteur avait pu être responsable de la distinction « *vivant / non vivant* » tout comme le HNR. La valeur de la centroïde spectrale change au cours du temps donc, pour chaque stimulus sonore, nous avons calculé les valeurs de la centroïde par trames de 1024 échantillons. De façon similaire au calcul du HNR, nous avons ensuite pris la valeur maximale pour chaque stimulus. Une ANOVA sur la catégorie a confirmé qu'il n'y a pas de différence significative entre les centroïdes spectrales des catégories « *vivante* » et « *non vivante* » (F(1,38)=4.03, $p = 0.052$). Ainsi le facteur de centroïde spectrale ne peut pas avoir été pris en compte par les participants pour effectuer cette distinction. Sur la figure 5.15, on se rend compte que les différentes données (maximum de Centroide Spectrale ; $p_{vivant}$) sont beaucoup plus éparses que dans le cas du HNR.

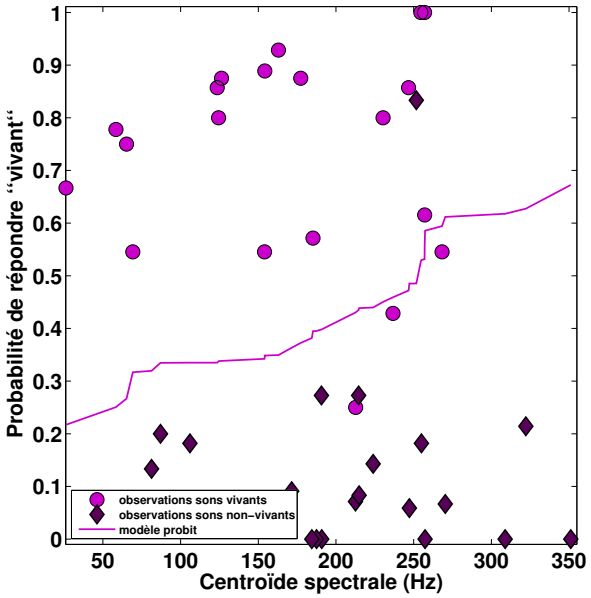



**Figure 5.15.** Mise en relation entre le choix des participants et la centroïde spectrale des stimuli pour une condition sans bruit ajouté (+max dB) et sans contexte visuel (Audio seul).

La courbe obtenue par un modèle probit est alors beaucoup plus plate, et beaucoup moins lisse, que celle obtenue pour le HNR. Cela indique qu'il n'est pas possible de mettre en relation les valeurs de centroïdes et les valeurs de $p_{vivant}$ par un modèle de type probit.

## 5.6 Discussion

### 5.6.1 Rôle du contexte

Les performances des participants en termes de taux d'identification correcte et de temps de réponse étaient améliorées lorsqu'un contexte visuel congruent était ajouté (**AVc**), tandis qu'un contexte visuel incongruent (**AVi**) dégradait ces performances. Ces résultats sont cohérents avec les études précédentes sur l'identification des sons d'environnement en présence d'un contexte visuel [Schneider *et al.* 2008; Ozcan et van Egmond 2009; Suied *et al.* 2007 2009].

Par ailleurs, contrairement à ce qui est observé pour des études où le contexte est une ambiance sonore, soit simultanée [Gygi et Shafiro 2011] soit en amorce [Gérard 2004], nous n'avons observé aucun avantage pour une condition incongruente.

### 5.6.2 Asymétre « vivant / non vivant »

Nous avons observé une différence significative de RT entre les deux catégories de sons, avec les sons de sources vivantes reconnus plus rapidement que ceux de sources non vivantes. Cette même différence en faveur des sons « *vivants* » était obtenue dans [Gérard 2004, p. 173] et ce quelque soit le type de contexte (amorce) employé (mot, phrase, ambiance sonore, son

d'environnement isolé). Cette différence s'observait aussi dans [Suied et Viaud-Delmon 2009], pour une présentation bimodale de la source, donc en présence d'une représentation visuelle ajoutée, tandis que dans notre étude cette différence est observable en présence ou non de représentation visuelle.

D'ailleurs, d'après l'étude de [Gérard 2004], l'effet du contexte dépend de la catégorie « *vivant / non vivant* ». Notamment, lorsque l'amorce était un son différent de la cible mais congruent sémantiquement, l'identification de la cible était facilitée pour les sons « *vivants* », mais retardée pour les sons « *non vivants* » pour un écart temporel court (200 ms) entre l'amorce et la cible. Dans notre étude, nous observons la même facilitation dans les deux catégories lorsque le contexte visuel est congruent.

La différence de mémorisation des sons issus de chaque catégorie peut partiellement expliquer la différence de temps de réponse observée entre les sons « *vivants* » et « *non vivants* ». En effet, une étude de [French et Mareschal 1998] laisse à penser qu'un plus grand nombre d'informations est stocké pour les items issus de la catégorie d'objets *manufacturés* (« *non vivants* ») que pour ceux de la catégorie *naturelle* (« *vivants* »).

Nous avons également observé une influence significative de la catégorie sur le taux de réponses correctes, avec les sons « *vivants* » mieux reconnus que les sons « *non vivants* » pour des RSB supérieurs à $-15$ dB, mais moins bien reconnus au-delà. Cette différence peut avoir été biaisée par le choix de la tâche. En effet, la chute soudaine du taux de réponses correctes pour les sons « *vivants* » (RSB $= -18$ dB) est accompagnée d'une augmentation inattendue du taux de réponses correctes pour les sons « *non vivants* ». Cette inversion brusque peut sous-entendre que les participants ont maintenu le nombre de réponses « *non vivantes* » mais ont seulement confondus les sons « *vivants* » avec des sons « *non vivants* ». Comme nous l'avons vu, cette confusion peut être due au HNR. Toutefois, pour pouvoir vraiment séparer l'influence de la catégorie sans faire intervenir le biais expérimental, nous pensons qu'une tâche n'impliquant pas de faire référence à la catégorie de la source aurait mieux convenu. Par exemple, [Schneider *et al.* 2008] proposent une autre tâche de catégorisation à deux possibilités : les participants doivent catégoriser en fonction de la taille de la source sonore à identifier. Ainsi, les sons « mouette » et « éléphant » auraient conduit à des réponses différentes bien que faisant partie de la même catégorie de sons « *vivants* ».

Enfin, ne cherchant pas ici à expliquer tous les facteurs responsables de cette différence entre catégories, nous notons surtout que la distinction « *vivant / non vivant* » ( ou *living* vs. *man-made*, ou encore *animate* vs. *inanimate*) n'est pas propre à l'identification de sources sonores. Sur le plan neurologique, des études montrent que les sons et les images produits par des animaux sont traités par des zones neuronales distinctes de celles qui traitent les sources inanimées [Lewis *et al.* 2005; Caramazza et Shelton 1998].

### 5.6.3 Indices acoustiques

L'analyse du rapport d'harmonicité (HNR) est en accord avec les études précédentes concernant l'identification de sons d'environnement [Giordano *et al.* 2010] et indique que ce paramètre influence le choix des participants : plus le HNR d'un son est élevé et plus les participants auront tendance à identifier ce son comme « *vivant* ». La probabilité d'utiliser ce facteur pour déterminer la catégorie du son a été calculée à travers un modèle de régression de type probit.

L'analyse de ces modèles a montré que les participants se sont fiés plus à l'harmonicité des sons lorsque le niveau de bruit était bas (RSB grand) et lorsqu'une image congruente était présentée en parallèle. Au contraire, les participants s'y sont moins fiés lorsqu'une image incongruente était ajoutée. Cela tend à indiquer que le traitement des informations acoustiques est influencé par le contexte visuel.

### 5.6.4 Utilisation des sons d'environnement en IHM

Les sons d'environnement sont utilisés dans les interfaces sonores [Gygi et Shafiro 2009], notamment sous forme d'icônes sonores [Brazil et Fernström 2011a]. Les résultats de notre étude soulignent plusieurs aspects qui pourront être utiles à la conception de telles interfaces.

La rapidité de traitement et la meilleure identification des sons « *vivants* », par rapport aux sons « *non vivants* », suggèrent qu'il est préférable d'utiliser cette catégorie de sons pour créer des icônes sonores plutôt que l'inverse. Il faudra cependant prendre en compte le rapport entre le son et ce qu'il représente (objet, animal ou évènement), car plus le rapport est direct et intuitif (lien sémantique fort), plus il est facile à mémoriser et à utiliser [Keller et Stevens 2004].

D'autre part, cette étude illustre l'effet de facilitation intersensorielle pour une présentation audiovisuelle congruente. Elle confirme ainsi l'avantage de la multimodalité, en particulier dans ce cas où l'une des deux modalités est dégradée. Toutefois, nous notons que les études sur le rôle du contexte sonore dans l'identification des sons d'environnement ont montré, au contraire, un avantage de l'incongruence entre le contexte sous forme d'ambiance sonore [Gérard 2004; Leech *et al.* 2009; Gygi et Shafiro 2011] et le son à identifier. Cet effet de *pop-out* du son improbable pourrait être judicieusement mis à profit dans une interface sonore pour indiquer un évènement de plus haute importance (un danger par exemple). Ainsi le concepteur d'une interface à base d'icônes pourrait se servir d'un ensemble de sons reliés au même environnement (comme ceux des scénarii de [Ballas et Mullins 1991]) pour indiquer des évènements habituels, et un ou deux sons incongruents avec cet environnement comme alarmes pour l'utilisateur. L'ensemble de sons « habituels » forme alors une palette sonore. Un exemple de telles palettes est utilisé dans une interface d'aide à la navigation pour les mal-voyants [Parseihian et Katz 2012a]. Toutefois, la notion d'incongruence sémantique pour référer aux évènements ou lieux dangereux (par exemple « une route à traverser » dans le cas de l'aide à la navigation) n'y est pas encore exploitée.

## 5.7 Perspectives

Dans cette étude nous n'étions concernée que par l'identification d'un seul son d'environnement à la fois. Le contexte et la présence d'autres sons simultanés étaient simulés par l'ajout d'un bruit blanc. Une étude plus écologique devrait s'intéresser à l'identification de plusieurs sons d'environnement simultanés et à l'apport d'un contexte sonore réaliste comme c'est le cas des études de [Gygi et Shafiro 2007a 2009] par exemple.

Par ailleurs, et de façon similaire à ce qui est utilisé communément dans d'autres études telles que [Suied *et al.* 2010; Ozcan et van Egmond 2009; Schneider *et al.* 2008], nos images sont statiques. Or nous avons vu en section 3.4.2 que le mouvement et la synchronisation audiovisuelle sont importants pour l'intégration des deux modalités. Par exemple, la lecture labiale lors de la

co-présentation d'une vidéo et d'une bande-son facilite la compréhension de la parole de cette bande-son. On peut alors se demander ce qu'il se serait passé si le contexte visuel avait été présenté dans l'expérience sous forme de vidéo.

Dans cette expérience notre but n'était pas de mesurer l'interaction audiovisuelle mais de mieux comprendre la perception des sons et en particulier des sons d'environnement. La tâche utilisée était donc d'identifier la source sonore et non l'objet audiovisuel. Notre étude évalue l'influence du visuel pour lever l'ambiguïté sonore. A ce stade si nous voulions mesurer l'intégration audiovisuelle pour : 1) évaluer le rôle de chaque modalité dans l'identification/la reconnaissance d'un objet ou 2) vérifier les prédictions d'intégration par les théories proposées dans la littérature, l'identification reposerait de plus en plus sur le visuel avec un niveau de bruit croissant (théorie MLE, optimisation bayésienne) et les RT seraient plus courts pour une présentation audiovisuelle que pour une présentation unimodale visuelle ou auditive seule. Pour répondre à ces questions, il faudrait un nouveau protocole expérimental qui mesure aussi l'identification dans une condition visuelle seule. Il faudrait alors modifier la tâche : les participants devraient juger la catégorie de la source audiovisuelle et non celle de la source produisant le son. On pourrait alors comparer les 3 modalités auditive, visuelle et audiovisuelle avec une association audiovisuelle congruente (pour éviter un conflit entre l'audio et le visuel où le participant ne pourrait se décider).

Chapitre 6

# Le flou audiovisuel : un critère perceptif pour augmenter la saillance d'objets multimédia ?

*Mes sens et ma conscience ne me livrent donc de la réalité qu'une simplification pratique.*
Henri Bergson, *Le Rire*.

## Sommaire

Comme nous l'avons vu dans l'état de l'art sur la perception (chapitre 3), certains paramètres augmentent la saillance perceptive d'un objet perçu auditivement ou visuellement. Notamment le flou est un paramètre visuel préattentif et peut être utilisé comme trait caractéristique pour guider l'attention vers une cible nette au milieu de distracteurs flous [Kosara *et al.* 2002b].

Dans ce chapitre, nous présentons une proposition pour étendre le flou visuel au domaine sonore et audiovisuel. Après avoir défini des flous audio et audiovisuel, une série de trois expériences a été mise en place pour évaluer : a) d'une part, l'influence des flous audio et audiovisuel sur une tâche de recherche, notamment l'éventuelle capture attentionnelle induite par ces paramètres ; b) d'autre part, le rôle respectif de chacune des deux modalités et de chacun des deux flous dans une tâche de recherche multimodale. Nous exposerons la méthodologie globale employée dans cette suite d'expérimentations, le corpus de stimuli créé pour l'étude, ainsi que les 3 expériences, respectivement visuelle, auditive et audiovisuelle.

Cette étude a été présentée en anglais lors de l'*International Multisensory Research Forum* (Oxford, juin 2012) sous le titre "*Redundancy Gains in Audiovisual Search*" et fait l'objet d'une publication en français pour la conférence *Ergo'IHM 2012* (Biarritz, octobre 2012) : "*Guidage attentionnel à base de flou audiovisuel pour la conception d'interfaces multimodales*". Ces deux publications étudient surtout la question de la redondance de la distorsion : est-il nécessaire de rendre flou chacune des deux modalités de manière cohérente, ou comme c'était le cas pour une lentille audiovisuelle grossissante (chapitre 4), est-ce qu'appliquer l'effet sur une seule des deux modalités suffit à améliorer la perception audiovisuelle dans des tâches de recherche multimodale ? La question de l'apport de la multimodalité a quant à elle donné lieu à une soumission pour le journal ACM Transactions on Applied Perception (soumission avril 2012, révision août 2012) sous le titre "Cueing Multimedia Search with Audio-Visual Blur".

## 6.1 Analogies de guidage : du flou visuel au flou audiovisuel

Il existe plusieurs types de flous visuels, notamment le ***flou statique*** (*static blur*) et le ***flou cinétique*** (*motion blur*). Peu utilisé tel quel dans le domaine de l'audio, le terme *flou* désigne souvent un manque de précision responsable d'une diminution de l'intelligibilité de la parole ou gênant l'identification des sons [Frederiksen 1967]. Il peut aussi être employé à des fins artistiques, en composition musicale notamment, comme dans les travaux de [Matsumoto 2009] ou [Keston 2009].

Nous nous proposons ici de trouver un équivalent audio au flou visuel avant de vérifier si ce

paramètre, que nous obtenons par analogie, induit les mêmes propriétés d'attraction qu'en visuel.

### 6.1.1 Le flou statique ou profondeur de champ

Le **flou statique** est associé à la profondeur de champ. Il est utilisé en IHM, notamment dans la technique de ***profondeur de champ sémantique*** [Kosara *et al.* 2002b; Kosara 2001; Schrammel *et al.* 2003; Rosenbaum et Schumann 2009] (voir la section 2.1.3.4, page 17). Ce type de flou est souvent utilisé en photographie car il permet d'attirer l'attention directement sur un objet net dans l'image. Il peut s'agir d'un portrait pour lequel on cherche à mettre le sujet en valeur. Le reste de la photographie, bien que renseignant sur le contexte, est alors de moindre intérêt et sera donc flouté. La figure 6.1 présente un exemple de ce type de photographie avec comme sujet une fleur.

**Figure 6.1.** Photographie d'une fleur. La mise au point sur la fleur au premier plan et la faible profondeur de champ permettent d'avoir la fleur nette devant un fond flou. Le focus attentionnel se porte alors sur la fleur.

Le flou statique visuel peut être obtenu par traitement d'image en filtrant les hautes fréquences spatiales. Il s'agit alors d'appliquer, par convolution, un filtre passe-bas 2D à l'image de départ. Les filtres les plus usités sont des filtres Gaussien définis par l'équation 6.1 :

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{6.1}$$

Dans cette équation, $G(x,y)$ représente la valeur du filtre aux coordonnées $x$ and $y$ ; et $\boldsymbol{\sigma}$, le *rayon* du flou, représente l'écart type. Plus le rayon $\boldsymbol{\sigma}$ est grand, plus l'image finale sera floue.

### 6.1.2 Analogie du flou statique en audio

Certains traitements du signal audio peuvent produire un effet de dégradation similaire au flou statique visuel. Notamment, par analogie algorithmique, on peut voir le filtrage des hautes

fréquences (HF) temporelles comme un effet de ***flou audio statique***. La perte de ces HF entraîne une dégradation où les détails sont moins présents : l'intelligibilité et l'identification des sons sont donc diminuées [Frederiksen 1967; Shafiro 2008b].

Pour obtenir cet effet de flou audio statique, on peut utiliser un filtre passe-bas à une dimension. La fréquence de coupure du filtre, notée $\boldsymbol{f_c}$, sera l'équivalent du rayon de flou visuel $\boldsymbol{\sigma}$ et l'on pourrait parler de rayon de flou sonore. Plus cette fréquence de coupure est basse, plus le signal audio filtré est flou. Un exemple d'un filtre de ce type est le filtre de Butterworth d'ordre $n$ dont la réponse en fréquence est représentée par le gain $G_n(f)$ dans l'équation 6.2 :

$$G_n(f) = |H_n(j \times 2\pi f)| = \left( \sqrt{1 + \left( \frac{f}{f_c} \right)^{2n}} \right)^{-1} \tag{6.2}$$

où $H_n$ représente la fonction de transfert du filtre, $j$ le nombre imaginaire, $n$ l'ordre du filtre, $f$ la fréquence, et $\boldsymbol{f_c}$ la fréquence de coupure.

### 6.1.3 Autre proposition : analogie du flou cinétique

Bien que l'étude qui suit ait porté exclusivement sur le flou statique tel que nous venons de le présenter dans les deux sections précédentes, nous nous sommes aussi intéressée à la notion de flou cinétique.

En photographie, le ***flou cinétique*** ou *motion blur* se produit lorsque le mouvement du sujet photographié est trop rapide par rapport au temps de pose. La photographie en figure 6.2 présente une personne immobile devant un train flou car en mouvement. On remarque que ce type de flou fait apparaître une trace sous forme de trainée (*streaking*) qui indique ainsi la direction du mouvement [1]. Par simplification on se sert souvent de traits pour indiquer un mouvement dans un dessin au crayon.

Nous avons pensé à deux analogies différentes pour représenter ce type de flou en audio. La première analogie situe les flous visuel et audio dans le domaine spatial. La perte de précision spatiale due au flou visuel peut être fortement rapprochée de la difficulté à localiser précisément un son dans l'espace. Ce manque de précision spatiale porte le nom de ***flou de localisation*** (*Auditory localization blur*) [Blauert 1997, p. 37].

La seconde analogie à laquelle nous pouvons penser pour étendre le flou cinétique à l'audio, passe du domaine spatial en visuel au domaine temporel en audio. La trainée visuelle qui s'étend d'un point à l'autre d'une image floue peut être comparée à une « trainée temporelle » en audio. Une manière d'obtenir un tel effet est d'ajouter de la réverbération au signal audio [Zölzer *et al.* 2002].

Plusieurs travaux sur l'intelligibilité de la parole en présence de plusieurs sources simultanées ont montré qu'elle est diminuée par la réverbération [Bradley *et al.* 1999; Shinn-Cunningham 2002; Culling *et al.* 2003]. Par ailleurs, la réverbération, et les réflexions acoustiques associées, biaisent les indices de localisation sonores ce qui entraîne une diminution de la ségrégation des différentes sources, essentielle pour l'attention sélective auditive [Bronkhorst 2000]. Ainsi, ces

1. Il ne faut cependant pas confondre le flou cinétique avec le flou de bougé qui lui est dû à un mouvement de l'appareil au moment de la prise de vue et non à un mouvement du sujet. Dans le cas d'un flou de bougé l'ensemble de la photographie est floue.

**Figure 6.2.** Photographie d'une personne immobile sur le quai devant un train en mouvement. Un effet de flou cinétique apparaît sur le train sous forme de trainées.

études confirment que la réverbération peut être considérée comme un effet de flou audio. Toutefois, nous n'avons pas connaissance d'études qui l'aurait utilisée comme trait caractéristique pour attirer l'attention sur une source sonore nette (non réverbérée). Une étude serait nécessaire pour vérifier que ce paramètre est aussi valable comme guide attentionnel d'un rendu audio que le flou visuel pour un rendu graphique.

### 6.1.4 Définition d'un flou audiovisuel

Maintenant que nous avons défini la notion de flou audio, nous pouvons penser à un flou audiovisuel pour flouter [2] de manière cohérente les deux composantes audio et visuelle des objets multimédia. Ainsi un flou audiovisuel sera obtenu en appliquant respectivement un flou visuel statique à la composante graphique de l'objet et un flou audio statique à la composante sonore de l'objet. Les niveaux de flou (par ex. $\boldsymbol{\sigma}$ pour un flou gaussien visuel et $\boldsymbol{f_c}$ pour un flou audio) seront alors indépendants pour les deux modalités. Notre travail s'est d'ailleurs en partie focalisé sur l'étude de la question suivante : comment régler le niveau de flou visuel et le niveau de flou audio pour obtenir des effets d'attraction du même ordre de grandeur dans les deux modalités pour ne pas biaiser la combinaison vers une des deux modalités ?

## 6.2 Méthodologie expérimentale

Nous présentons dans cette section, d'une part, les objectifs et les différents points évalués dans cette étude, et d'autre part la démarche expérimentale que nous avons suivie pour les évaluer.

2. Bien que le terme « flouter » soit un néologisme, il est couramment employé en traitement d'images pour désigner l'action de rendre floue une image ou une vidéo.

### 6.2.1 Objectifs

Cette étude avait pour but d'évaluer plusieurs aspects :

- **L'efficacité du flou comme facilitateur de recherche** : il s'agit de comparer les effets du flou (réciproquement de la netteté) quand il est appliqué sur la cible ou sur les distracteurs. Cela permet de vérifier que le flou audio et le flou audiovisuel permettent les mêmes effets d'attraction ou capture attentionnelle que le flou visuel ;
- **Le rôle de chaque modalité et l'effet de l'intégration multimodale** : il s'agit de confirmer le gain d'une présentation audiovisuelle par rapport à une présentation unimodale, c'est-à-dire l'*effet de cibles redondantes* tel que l'état de l'art le suggère. Nous comparerons donc les résultats obtenus avec la combinaison bimodale à ceux obtenus dans chacune des deux recherches unimodales. Cependant, il y a un risque d'avantager une des deux modalités plus que l'autre par le choix d'un niveau de flou qui induirait un effet d'attraction plus important dans une modalité que dans l'autre. Pour limiter ce risque, nous procéderons d'abord à une phase de ***calibration multimodale*** par laquelle nous sélectionnerons un niveau de flou dans chaque modalité afin d'obtenir des performances du même ordre de grandeur lors d'une recherche audio et d'une recherche visuelle.
- **La contribution de chacun des flous visuel et auditif dans l'influence du flou audiovisuel** : il s'agit de mesurer l'effet de la redondance d'un effet de dégradation des distracteurs, c'est-à-dire la répétition d'un paramètre facilitateur (ou trait caractéristique), dans deux modalités différentes.

### 6.2.2 Approche générale

Afin de comparer la contribution de chaque modalité dans la recherche audiovisuelle, il fallait d'abord ajuster l'effet d'attirance du trait caractéristique dans chaque tâche de recherche unimodale. Nous avons donc réalisé une étape de ***calibration intermodale***. En effet, ajuster le niveau de flou, c'est-à-dire régler le rapport netteté/flou, est nécessaire pour que la netteté agisse comme un trait caractéristique [Kosara 2001; Schrammel *et al.* 2003, p. 32] : si la différence de netteté entre l'objet net et les objets flous n'est pas suffisante, l'objet net n'attirera pas nécessairement l'attention. Au contraire, si le flou est trop important, les objets flous deviennent inutiles car incompréhensibles. Cette étape de calibration est nécessaire car il y a un risque de choisir un niveau de flou plus fort dans une des deux modalités, entraînant alors un effet d'attraction plus important et des temps de réponse (RT) plus courts pour cette modalité. Ce genre de calibration multimodale a d'ailleurs été précédemment utilisé par [Schneider *et al.* 2008] dans une expérience de reconnaissance d'objets sonores et visuels. Les stimuli visuels étant initialement mieux reconnus que les stimuli sonores, ils étaient dégradés par ajout de bruit pour les tests audiovisuels, de sorte que les taux d'identification soient similaires pour les stimuli visuels et sonores pris individuellement.

L'étude que nous avons menée a alors été séparée en trois expériences distinctes. Les deux premières expériences évaluaient l'influence de différents niveaux de flous dans des tâches de recherche unimodales. Ainsi l'expérience 1 comparait différents niveaux de flous visuels sur une tâche de recherche visuelle et l'expérience 2 comparait différents niveaux de flous audio sur une tâche de recherche auditive. En comparant les résultats de ces deux expériences nous avons

sélectionné un niveau de flou visuel et un niveau de flou auditif conduisant à des performances comparables en termes de RT et d'identification. Ces niveaux de flou ont ensuite été utilisés durant l'expérience 3. Cette expérience avait pour but d'évaluer le rôle de chaque modalité dans une présentation audiovisuelle. Elle comparait donc les performances d'une tâche de recherche bimodale avec les performances unimodales. L'expérience 3 suivait le même paradigme expérimental que les expériences 1 et 2, c'est-à-dire la même tâche, les mêmes conditions de congruence et les mêmes stimuli (soit avec chaque modalité indépendante ou en association). Ce paradigme est détaillé ci-après.

### 6.2.3 Hypothèses

Compte tenu des études précédentes sur la recherche unimodale et l'intégration multimodale, on peut émettre plusieurs hypothèses.

Hypothèse 1 **Effet de dégradation** : comme le flou est avant tout une dégradation, l'identification d'une cible floue devrait induire des performances moindres que pour l'identification d'une cible nette, quelque soit l'état des distracteurs. Les taux de réponses correctes devraient être inférieurs et RT plus longs pour toutes les modalités. Cette hypothèse a été testée dans l'expérience 1 pour le visuel, dans l'expérience 2 pour l'audio et dans l'expérience 3 pour l'ensemble des modalités.

Hypothèse 2 **Avantage de l'effet de pop-out (facilitation) en cas de cohérence entre le trait caractéristique et la cible** : une cible nette devrait attirer l'attention si les distracteurs sont flous, c'est-à-dire qu'elle devrait être retrouvée plus rapidement si les distracteurs sont flous que s'ils sont nets. Comme pour l'Hypothèse 1, cette hypothèse a été testée dans les trois expériences.

Hypothèse 3 **Désavantage de l'effet de pop-out ou capture attentionnelle (complication) en cas d'incohérence entre le trait caractéristique et la cible** : un distracteur net devrait aussi attirer l'attention même s'il n'est pas pertinent pour réaliser la tâche. Repérer et identifier la cible floue parmi ce distracteur net et les autres flous devrait donc être plus long que dans une condition où tous les objets sont flous, puisqu'il faut d'abord inhiber l'attirance vers le distracteur net. Comme pour les Hypothèses 1 et 2, cette hypothèse a été testée dans les trois expériences.

Hypothèse 4 **Effet de cibles redondantes** : une combinaison cohérente[3] des modalités visuelles et auditives devrait conduire à de meilleures performances que celles obtenues pour chacune des modalités prises individuellement. Cette hypothèse a été testée dans l'expérience 3.

Hypothèse 5 **Redondance de traits caractéristiques** : dans un cas de congruence entre la cible et le trait caractéristique, la combinaison de traits caractéristiques auditif et visuel devrait améliorer les performances par rapport à l'utilisation d'un trait caractéristique unimodal. Les performances devraient être meilleures si les

3. Nous employons ici le terme « cohérent » pour indiquer une présentation audiovisuelle synchronisée temporellement et spatialement et où les données des deux modalités se correspondent sémantiquement.

distracteurs sont flous sur leur deux composantes, auditive et visuelle, et pas uniquement sur une des deux composantes. Cette hypothèse a également été testée dans l'expérience 3.

### 6.2.4 Tâche et conditions expérimentales

Durant chaque essai, six stimuli (cinq distracteurs et une cible) étaient présentés simultanément. La cible pouvait être l'une de deux options et les participants devaient identifier laquelle de ces deux options possibles était présente dans l'ensemble des stimuli. Il s'agissait donc d'une tâche de type *two-alternative forced choice* (2AFC). En pratique, et comme cela est justifié à la section 6.3, nous avons utilisé pour cibles les nombres 6 et 10. Une et une seule des deux options de cible était présente. Les participants devaient donc répondre à la question « 6 ou 10 ? ». Ils répondaient en pressant une des touches sur le clavier (des marques « 6 » et « 10 » étaient indiquées sur les touches 6 et 0). L'assignation des touches était contrebalancée pour la moitié des participants.

Les stimuli étaient présentés soit visuellement (**V**) (expériences 1 et 3), soit auditivement (**A**) (expériences 2 et 3), soit par une combinaison audiovisuelle, le flou pouvant alors être appliqué sur l'une, l'autre (**AV**$^{\mathbf{neut}}$ ou **A**$^{\mathbf{neut}}$**V**) ou les deux modalités à la fois (**AV**) (expérience 3).

Habituellement les études sur la préattention visuelle déterminent si un paramètre peut être considéré comme préattentif ou non en évaluant l'influence du nombre de distracteurs sur le temps de recherche. Dans notre cas, nous n'avons pas utilisé cette méthode de détermination du fait de l'usage de l'audio. La présence de trop nombreuses sources simultanées aurait rendu la tâche trop difficile voire impossible à réaliser. De plus, cette tâche réalisée sur un trop grand nombre de sources ne présente que peu d'intérêt si l'on considère qu'à long terme il s'agit de présenter une sélection de documents issus des résultats d'une recherche par mots clés. Pour évaluer si notre paramètre de flou pouvait fonctionner comme un guide attentionnel involontaire et sans effort, nous avons donc proposé un autre protocole expérimental, original, où il s'agissait de comparer les performances obtenues sous quatre conditions de congruence différentes entre le trait caractéristique et la cible.

Si le trait caractéristique (netteté) est appliqué à la cible, il s'agit d'une condition **Congruente**, s'il est appliqué à un distracteur mais que la cible reste floue, il s'agit d'une condition **Incongruente**. Deux autres conditions de congruence ont été testées pour servir de « contrôle » : une condition **Neutre**, dans laquelle les six stimuli présentés en même temps (cible + 5 distracteurs) étaient nets, et une condition **Dégradée** dans laquelle les six stimuli étaient floutés avec le même niveau de flou. Les quatre conditions sont présentes dans chacune des trois expériences.

Pour clarifier et faciliter la mémorisation du protocole expérimental, nous avons utilisé, dans tous les graphiques, un même code couleur pour faire référence à ces conditions de congruence. Ainsi la condition **Neutre** est représentée par du rouge, la condition **Congruente** par du bleu, la condition **Incongruente** par du vert, et la condition **Dégradée** par du noir.

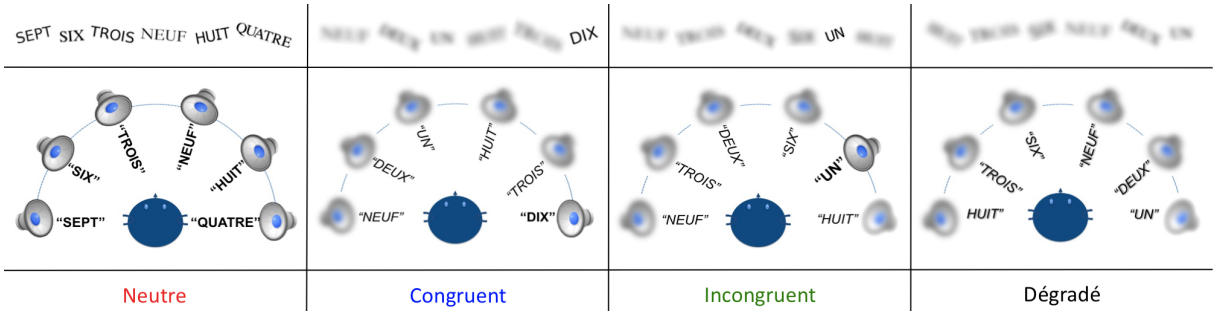



**Figure 6.3.** Représentation schématique de la position et de l'état de chaque stimulus dans une présentation visuelle (haut, vue de face) ou audio (bas, vue de dessus) pour les différentes conditions de congruence. La cible recherchée est soit le SIX, soit le DIX.

### 6.2.5 Cohérence spatiale et temporelle entre présentation audio et présentation visuelle

Dans les cas visuel seul ou audiovisuels, les six stimuli étaient présentés dans un ordre aléatoire, alignés, uniformément espacés et orientés aléatoirement entre $-45°$ et $+45°$ (voir ligne du dessus de la figure 6.3). Les stimuli visuels apparaissaient en noir sur un fond gris à 50 %.

Comme la ségrégation spatiale est un paramètre facilitateur dans les tâches de recherche auditive [Eramudugolla *et al.* 2008], nous avons spatialisé les stimuli audio pour aider la ségrégation des différents flux [Bronkhorst 2000; Bregman 1990; Brungart et Simpson 2005]. Pour spatialiser les sources, nous avons utilisé un procédé de synthèse binaurale non individualisée qui permet d'obtenir un rendu sur casque, respectant ainsi notre volonté d'accessibilité future au grand public. Pour cette expérience nous avons utilisé le LSE [Katz *et al.* 2010] comme moteur de spatialisation. Puisque les stimuli visuels étaient présentés en ligne, nous avons gardé une certaine cohérence spatiale et les sons sont spatialisés virtuellement le long d'un arc de cercle horizontal aux azimuts $\pm 18°$, $\pm 54°$ et $\pm 90°$ (voir la ligne du bas de la figure 6.3). Pour les essais bimodaux, l'ordre des stimuli audio était le même que celui des stimuli visuels et l'on peut donc dire que les stimuli audio et visuels étaient localisés de manière cohérente bien qu'ils ne soient pas coïncidents.

D'un point de vue temporel, les stimuli visuels restaient fixes tandis que les stimuli audio étaient joués en boucle jusqu'à ce que le participant réponde. Le bouclage des stimuli audio était fait de sorte que les stimuli audio se répètent toujours avec un départ (*onset*) synchronisé pour les six stimuli : la durée de la boucle correspondait donc à la durée du plus long stimulus de l'essai comme représenté par les formes d'ondes des signaux audio en figure 6.4.

## 6.3 Étude des stimuli

Comme pour le réglage du niveau de flou que nous présentons par la suite (calibration multimodale), les expérimentations doivent porter sur des stimuli audio et visuels qui ne favorisent pas l'une des deux modalités. En particulier, les stimuli d'une modalité ne doivent pas deman-

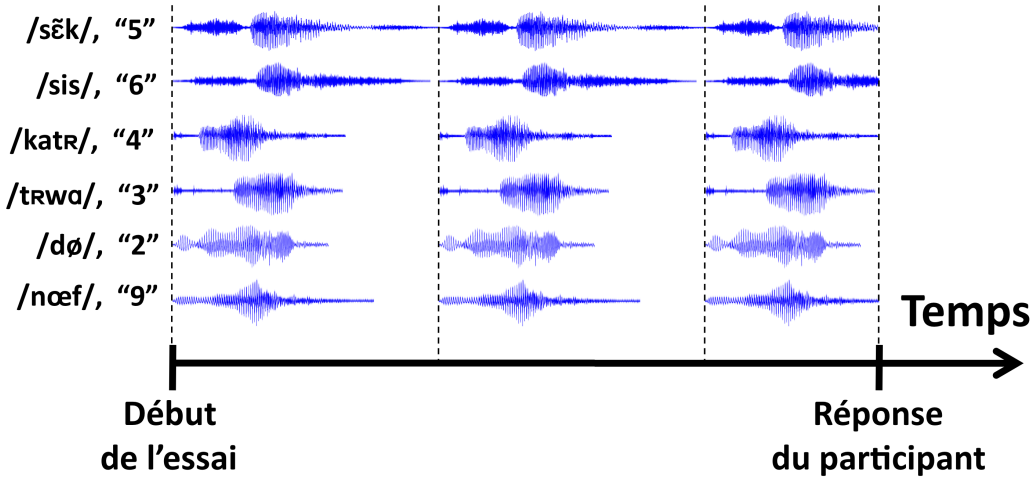



**Figure 6.4.** Les stimuli audio sont lus en boucle calés sur le stimuli le plus long (ici le son /$s\tilde{\varepsilon}k$/) jusqu'à ce que le participant réponde.

der une charge cognitive plus importante que ceux de l'autre modalité. De plus, nous avons besoin qu'il y ait une certaine congruence sémantique entre le contenu audio et visuel de chaque stimulus audiovisuel, de façon à ce que les contenus soient facilement associés et qu'une intégration multisensorielle soit possible. Parmi les études sur la recherche visuelle, Schrammel et ses collègues [Schrammel *et al.* 2003] ont utilisé des ellipses auxquelles il est difficile d'associer une composante sonore. D'autres études ont utilisé des lettres, par exemple dans [Kwak *et al.* 1991] la tâche consistait à retrouver la lettre « T » parmi plusieurs lettres « L ». Il serait plus simple de trouver une composante sonore à associer à ces stimuli visuels sous forme de lettres compte tenu du fort lien entre le langage écrit et le langage parlé. Au final, nous avons choisi de travailler avec des nombres comme c'est le cas dans des études sur la recherche visuelle comme [Krummenacher *et al.* 2001], mais aussi dans des études sur la perception sonore de la parole en présence de locuteurs multiples comme dans [Brungart et Simpson 2005] où les auteurs utilisent un système anglophone de coordonnées, de noms et de nombres.

Les nombres semblaient convenir car ils ont, grâce au langage, une correspondance naturelle entre audio et visuel. Cependant, puisqu'il y a plusieurs façons de représenter des nombres, langage ou répétitions en audio, dénombrement, mots ou chiffres en visuel, il restait encore à déterminer sous quelle forme exacte nous allions présenter ces nombres. Enfin, il fallait vérifier la validité des stimuli sélectionnés pour une étude audiovisuelle.

### 6.3.1 Proposition de stimuli visuels

Plusieurs propositions de stimuli visuels ont été nécessaires avant d'aboutir à un ensemble de stimuli cohérents pour l'expérimentation. Chacune des propositions était testée individuellement en expérience pilote avec trois à six participants. La tâche était celle décrite en section 6.2.4, p. 128. Nous ne détaillerons pas les résultats mais nous présentons les raisons majeures pour lesquelles les stimuli que nous avons rejetés ne nous ont pas semblé convenir.

#### 6.3.1.1 Etoiles à n branches

Notre première idée a été d'utiliser des étoiles à $n$ branches, $n$ variant de 1 à 10. La tâche consistait alors à déterminer si une étoile à 6 branches était présente où s'il s'agissait d'une étoile à 10 branches. La figure 6.5 présente deux exemplaires d'étoiles à 6 branches et deux exemplaires d'étoiles à 10 branches. Les stimuli visuels étaient alors faciles à confondre cas assez similaires dans la forme.

Après l'expérience pilote, il s'est avéré que nous avions omis un défaut de cette technique : au-delà d'un nombre de branches assez petit (au-delà de 3), il faut compter les branches une par une et ce pour chaque étoile qui a trop de branches. Ce traitement de dénombrement n'est pas préattentif et requiert un laps de temps supplémentaire. Le gain de temps de l'attraction d'une cible nette était alors très faible en comparaison du temps nécessaire au comptage du nombre de branches de la cible nette.


**Figure 6.5.** Proposition de stimuli visuels en forme d'étoiles à 6 et 10 branches.

#### 6.3.1.2 Nombres sous forme écrite

Pour éviter de forcer le participant à dénombrer les éléments constitutifs de chaque stimulus, nous avons décidé de représenter les nombres de 1 à 10 directement sous forme de chiffres arabes. C'est une remarque des participants qui nous a fait prendre conscience du problème réel de cette écriture en chiffres : tandis que le nombre 10 est formé de deux chiffres, le nombre 6 n'en contient qu'un. La distinction était alors trop facile entre les deux cibles 6 et 10 (voir figure 6.6). Les performances relevées lors de l'expérience pilote montraient d'ailleurs un taux de reconnaissance quasi parfait même lorsque la cible était très floutée.

En écrivant les nombres en lettres minuscules plutôt qu'en chiffres, nous réduisions la trop grande facilité de distinction entre le 6 et le 10 dans un cas flou. Toutefois, la barre de la lettre "d" dépassant du reste des lettres par rapport au "s", il était encore trop facile de distinguer le 6 du 10, même avec un fort niveau de flou.

Nous avons donc choisi d'écrire les nombres en lettres majuscules. Ainsi, une fois floutée, les lettres "S" et "D" se confondent. Les stimuli SIX et DIX sont donc très proches. Nous avons vérifié, par des tests objectifs présentés dans la section 6.3.3, que le choix de stimuli sous forme de mots écrits en toutes lettres était justifié avant de poursuivre.

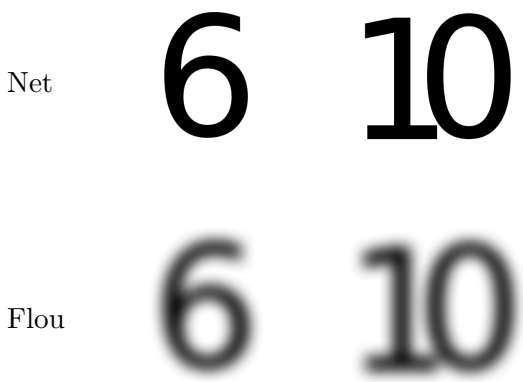



**Figure 6.6.** Proposition de stimuli visuels écrits en chiffres arables. Le 10 est trop distinct du 6, car il est formé de deux chiffres.

#### 6.3.1.3 Stimuli visuels sélectionnés

Chaque stimulus visuel était un nombre écrit en lettres capitales. Pour créer ces stimuli, nous avons écrit chaque nombre sous quatre exemplaires différents en utilisant quatre polices de caractère (entre parenthèses, la taille de la police en pixels) : Sans Serif(60), Times(67), Tahoma(60), STSong(68). La taille de chaque police était adaptée de sorte que la hauteur des lettres soit identique pour toutes les polices. Les mots étaient centrés sur des images de 267×267 px (pixels). Par la suite les stimuli étaient présentés sur un écran 19" de résolution 1280×1024 px placé à une distance d'environ 60 cm. Les stimuli visuels s'étendaient donc sur un angle visuel allant de 2.1° pour le mot "UN" à 3.0° pour "QUATRE".

### 6.3.2 Proposition de stimuli sonores

Les stimuli audio étaient des enregistrements des nombres de 1 à 10 prononcés en français par une voix féminine. Chaque nombre était enregistré quatre fois pour obtenir quatre versions distinctes du même mot (légères différences d'intonation et de durée par exemple). Tous les stimuli audio étaient cependant prononcés par la même voix pour garder une cohérence sur la largeur de bande spectrale utile. Les stimuli étaient enregistrés à une fréquence d'échantillonnage de 44100 Hz sur 16 bits. La durée moyenne des stimuli était de 456 ms (min 204 ms – max 685 ms).

Ces stimuli sonores ayant donné de bons résultats en expérience pilote, ce sont ceux que nous avons utilisés pour l'expérience 2 et l'expérience 3.

### 6.3.3 Comparaison objective des stimuli : mesures de similarité

Avant de procéder à l'expérience, nous avons décidé de vérifier que nos stimuli, écrits en majuscule ou prononcés, étaient suffisamment bien choisis d'un point de vue objectif. En particulier, nous voulions nous assurer que le 6 et le 10 étaient suffisamment similaires, tout en étant suffisamment éloignés des autres nombres, pour qu'il n'y ait pas d'ambiguïté possible entre une

cible et un distracteur. Pour cela nous avons calculé des mesures de similarité objectives en calculant la distance entre chaque paire de stimuli visuels puis entre chaque paire de stimuli audio.

#### 6.3.3.1 Similarité visuelle : distance couleur CIELab

Comme mesure de similarité entre deux images, nous avons utilisé la distance couleur CIE-Lab $\triangle E_{ab}$, qui est une distance euclidienne dans cet espace couleur, telle que définie dans [Chareyron 2005, p. 62] :

$$\triangle E_{ab} = \sqrt{(L_2 - L_1)^2 + (a_2 - a_1)^2 + (b_2 - b_1)^2} \tag{6.3}$$

où $L1$, $a1$ et $b1$ sont les trois valeurs décrivant la première couleur dans l'espace $L * a * b$ (avec L la luminance, et $a$ et $b$ les valeurs chromatiques respectivement pour l'axe rouge-vert et l'axe bleu-jaune), et $L2$, $a2$ et $b2$ les trois valeurs décrivant la seconde couleur. On rappelle que contrairement au système de couleur RVB, habituellement utilisé pour l'acquisition et le stockage des images numériques, car couvrant bien le spectre des couleurs visibles en photographie et étant assez simple à interpoler, le système couleur $L*a*b$ (ou CIELab) est un système de couleur dit perceptuel car il cherche à imiter la réponse de l'œil humain. Il a été conçu spécialement pour que la distance entre deux couleurs dans cet espace représente la perceptibilité de la différence entre ces deux couleurs. Ainsi d'après l'équation 6.3, si $\triangle E_{ab}$ est inférieur à 1 les deux couleurs sont perceptivement identiques. Cependant, dans notre cas, les images étant en noir et blanc, la mesure de distance euclidienne (comparaison pixel par pixel) aurait pu s'effectuer avec n'importe quel système de couleur.

La distance entre deux images $\overline{\triangle E_{ab}}$ s'obtient en calculant une distance couleur moyenne. Cette distance est définie comme la moyenne sur l'ensemble des pixels de la distance couleur des pixels de même position dans les deux images [Chareyron 2005, p.112] :

$$\overline{\triangle E_{ab}} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (\overrightarrow{Im1_{Lab}}(i,j) - \overrightarrow{Im2_{Lab}}(i,j))^2 \tag{6.4}$$

Nous notons que cette mesure de similarité ne prend pas en compte les éventuels décalages dans l'espace, les rotations ou toute autre transformation qu'il peut y avoir entre deux images. Ainsi deux images peuvent être mesurées comme très dissimilaires avec cette mesure, là où un individu les aurait correctement appariées. Nous nous attendons donc à des différences entre deux exemplaires/polices d'un même mot. Cependant comme tous nos stimuli visuels sont centrés et alignés, cette distance moyenne devrait nous donner des mesures de similarité suffisantes pour vérifier notre hypothèse sur la similarité entre le 6 et le 10.

#### 6.3.3.2 Similarité visuelle : comparaison inter-stimuli et effet du niveau de flou.

En comparant les stimuli par paire, on obtient la matrice de dissimilarité figure 6.7. Sur cette matrice on observe les distances entre chaque paire de stimuli. Plus la similarité est grande et plus la distance est faible (bleue).

Pour plus de lisibilité et une meilleure interprétation, nous avons également effectué un calcul de positionnement multidimensionnel (*Multidimensional Scaling* ou MDS). Nous avons gardé

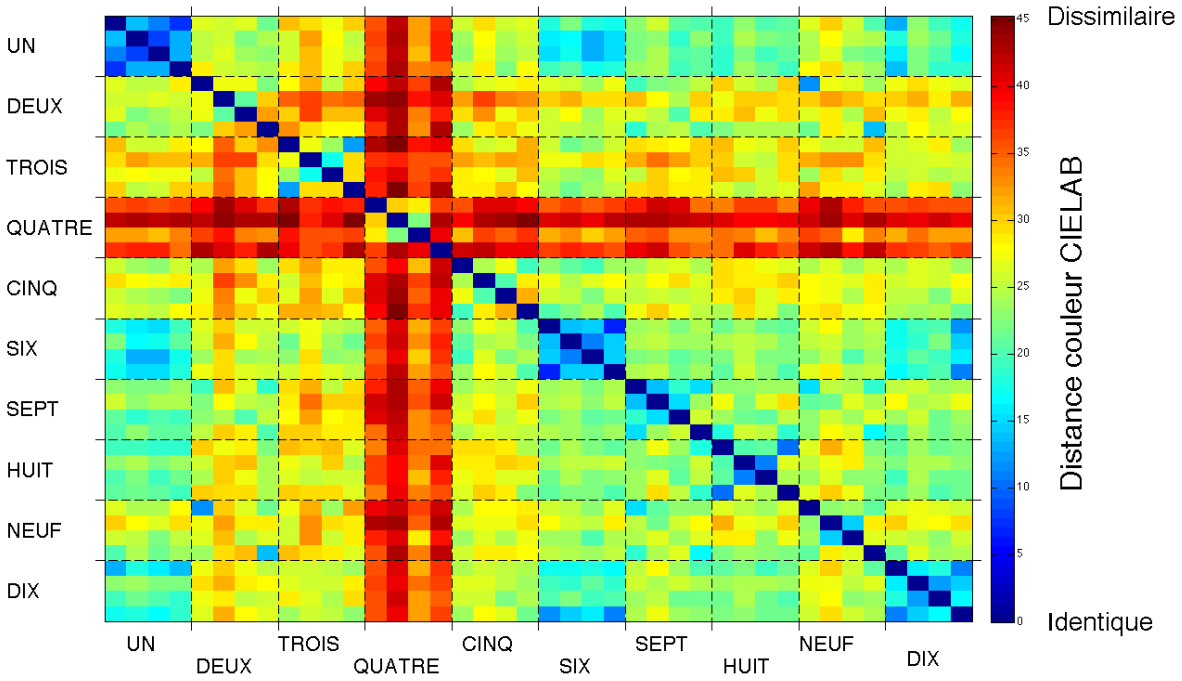



**Figure 6.7.** Matrice de dissimilarité sur les stimuli visuels originaux. Obtenue à partir d'une distance moyenne sur la distance couleur CIELab.

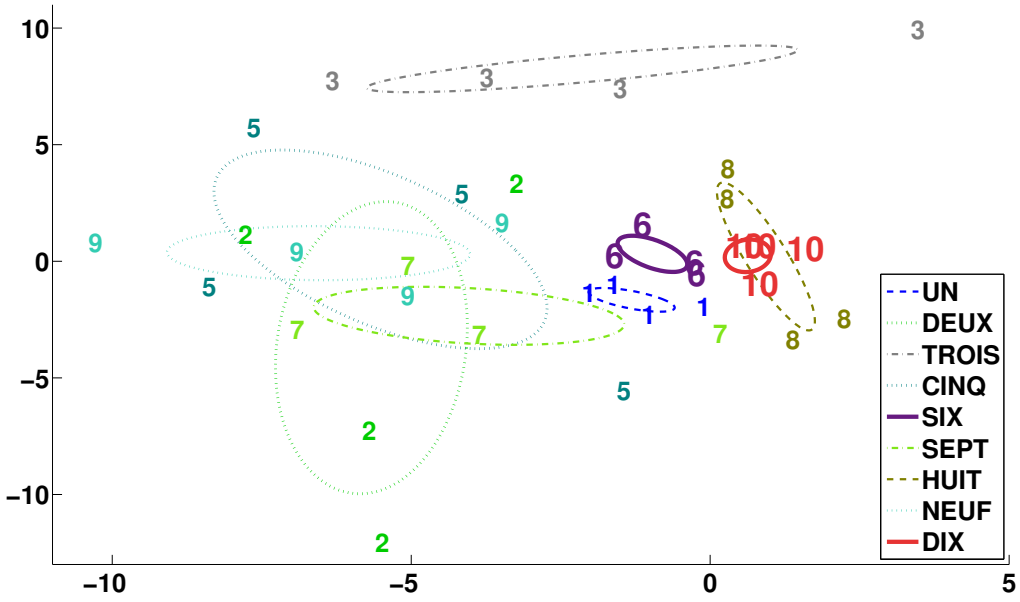



**Figure 6.8.** Projection des stimuli visuels sur l'espace 2D formé par les deux premiers axes du MDS obtenu sur la matrice de distance couleur CIELab figure 6.7. Chaque nombre représente un exemplaire. Les exemplaires du nombre 4 ne sont pas représentés parce qu'ils sont situés trop loin des autres (centre de l'ellipse : (24.6,−2.2) ; écart sur x : 17.9 – 33.6 ; écart sur y : −23.7 – 10.7). Les ellipses représentent des ellipses de confiance.

les deux composantes principales de cet MDS afin de représenter les stimuli dans un espace 2D en figure 6.8.

Nous observons alors que les ellipses associées aux exemplaires des SIX et des DIX sont très proches dans cet espace. Ces nombres sont donc bien choisis pour être deux options possibles de la cible, car ils seront effectivement plus facilement confondus que les autres paires. Le UN et le HUIT semblent toutefois assez proches. On peut supposer qu'une fois un flou appliqué sur le SIX ou le DIX, ils puissent non seulement être confondus mais également être confondus avec le UN ou le HUIT. On remarquera également la grande dissimilarité entre les stimuli QUATRE et le reste des stimuli. Cela peut s'expliquer par la longueur plus importante du mot QUATRE (6 caractères) qui occoupe alors plus de pixels que les autres mots qui sont constitués en moyenne de 4 caractères seulement.

La figure 6.9 présente l'influence du flou visuel sur les stimuli, en particulier sur les options de cible potentielles. Plus le flou augmente, plus les deux options de cible semblent difficiles à distinguer.

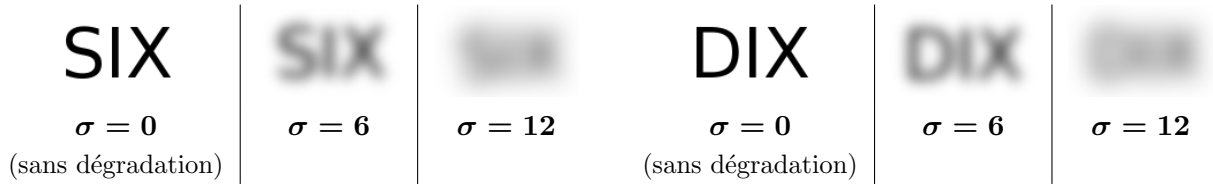



**Figure 6.9.** Effet du flou visuel (filtrage passe-bas gaussien) sur les options de cibles potentielles pour différentes valeurs de rayon $\sigma$.

Une autre matrice de dissimilarité, comparant les distances couleur moyennes CIELab telles que définies dans l'équation 6.4 pour les deux options de cible et différents niveaux de flou, est présentée figure 6.10. Le flou augmente la similarité entre deux stimuli. Cela appuie notre Hypothèse 1 selon laquelle l'identification d'une cible floue est plus difficile que celle d'une cible nette et donc que le flou peut être vu comme une dégradation perceptive.

#### 6.3.3.3 Similarité audio : mesures pour la reconnaissance automatique de la parole

Pour comparer les stimuli audio, nous avons utilisé une mesure de distance employée en reconnaissance automatique de la parole [Mermelstein 1976]. Il s'agit d'abord de calculer 13 coefficients cepstraux (MFCC pour *Mel Frequency Cepstral Coefficients*) par trames de 256 échantillons pour chacun des stimuli. Cela permet d'analyser le contenu fréquentiel sur chacune des trames. Ensuite, pour chaque paire de stimuli à comparer, un alignement temporel est effectué sur les MFCC grâce à un algorithme appelé DTW (*Dynamic Time Warping*). Cet alignement temporel réduit la distance entre deux mots si elle est due à une vitesse de prononciation différente. Une description complète de l'algorithme DTW est disponible dans [Sakoe et Chiba 1978]. Nous en présentons ici les étapes importantes :

1. Chaque trame temporelle du premier stimulus est comparée à chaque trame du second stimulus en calculant une distance euclidienne ($\triangle TrameToTrame$) sur les 13 MFCC $C_i$

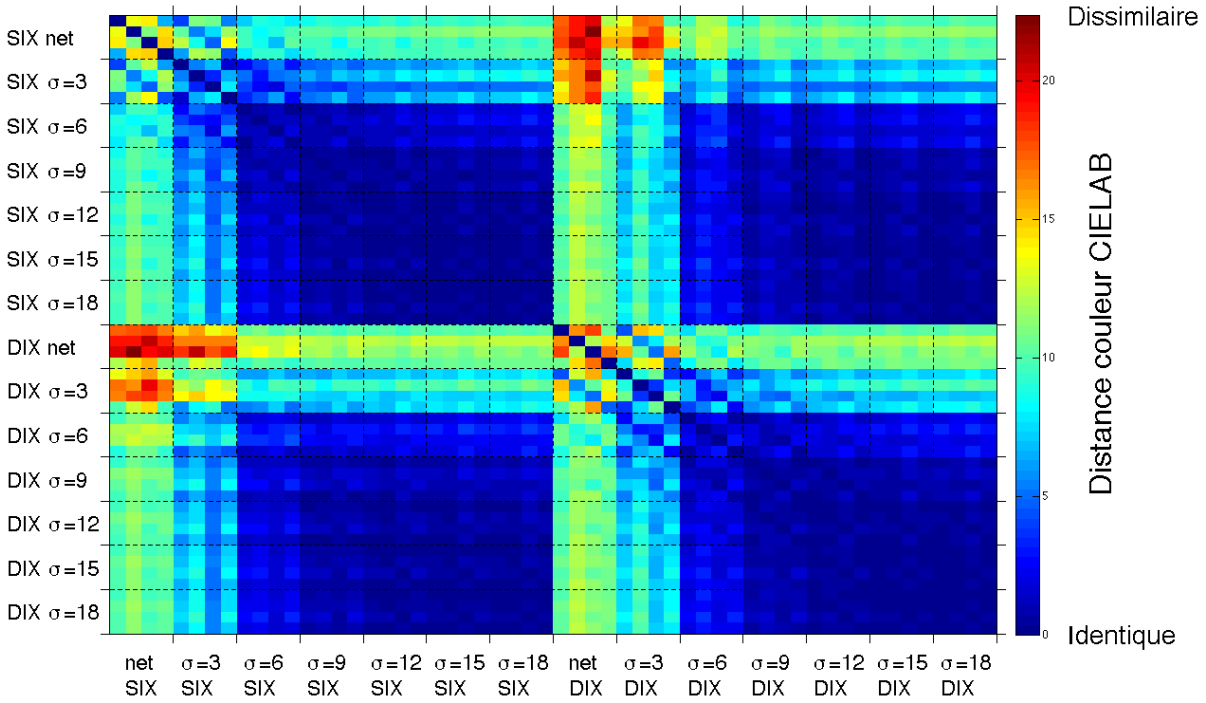



**Figure 6.10.** Matrice de dissimilarité pour les différentes versions, originales ou floues, des stimuli "cibles" 6 et 10.

(Equation (4)).

$$\triangle TrameToTrame = \sqrt{\sum_{i=1}^{13}(C_{i1} - C_{i2})} \tag{6.5}$$

2. La distance globale entre les deux stimuli est alors obtenue en minimisant le coût d'alignement (on minimise la somme des $\triangle TrameToTrame$).

#### 6.3.3.4 Similarité audio : comparaison inter-stimuli

Comme pour les stimuli visuels, nous avons effectué un calcul de positionnement multidimensionnel (MDS) sur la matrice de similarité des stimuli sonores (figure 6.11). La figure 6.12 présente la position des différents stimuli sonores dans un espace 2D formé par les deux composantes principales de l'analyse MDS.

Comme pour le visuel, nous observons que les ellipses associées aux exemplaires des sons /*sis*/ et /*dis*/ sont très proches dans cet espace. Ces nombres sont donc bien choisis pour être deux options possibles de la cible, car ils seront effectivement plus facilement confondus l'un avec l'autre qu'avec d'autres stimuli.

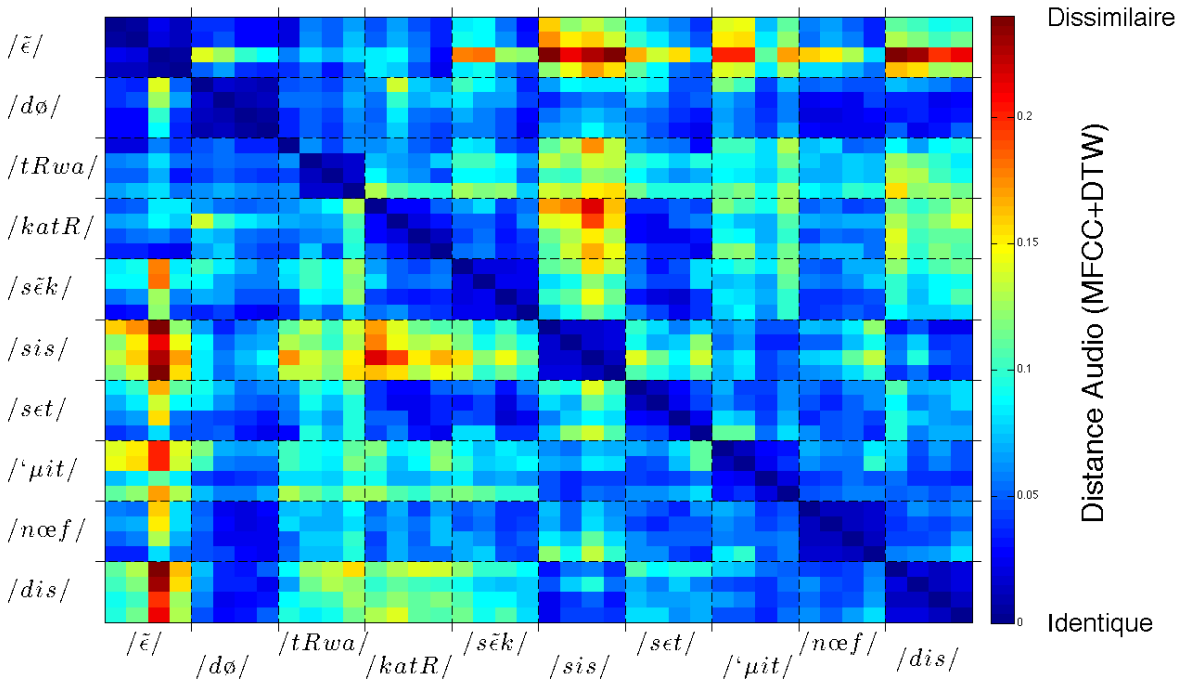



**Figure 6.11.** Matrice de dissimilarité sur les stimuli sonores originaux. Distances obtenues par l'algorithme de DTW.

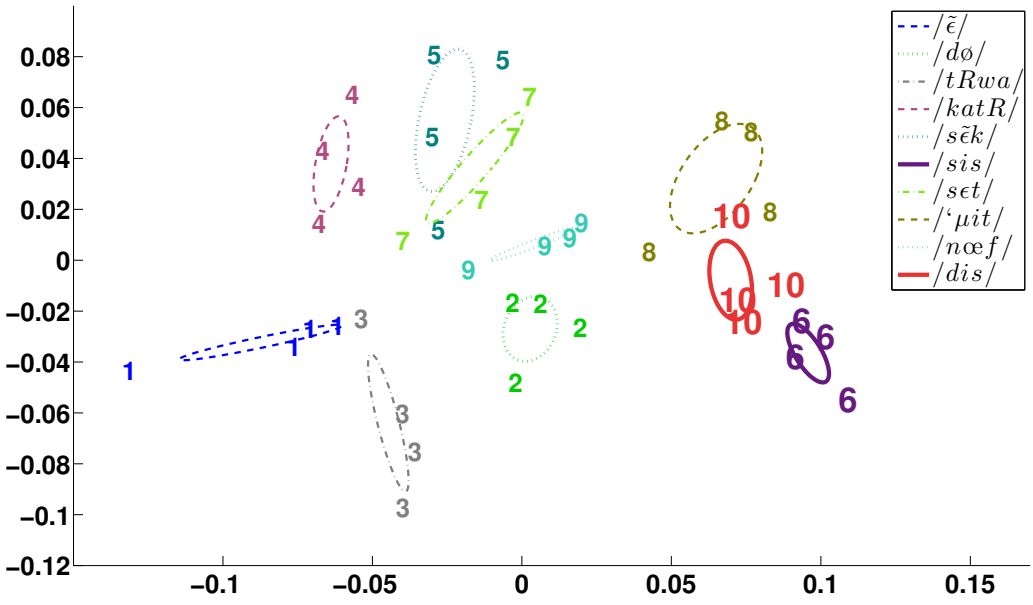



**Figure 6.12.** Projection des stimuli auditifs sur l'espace 2D formé par les deux premiers axes du MDS obtenu sur la matrice de distance. Chaque nombre représente un exemplaire du son prononcé. Les ellipses représentent des ellipses de confiance.

## 6.4 Expérience 1 : Effet du niveau de flou visuel sur une recherche visuelle

Les études disponibles sur les tâches de recherche visuelle, en particulier celles qui étudient l'effet du flou et la netteté comme trait caractéristique, utilisent des ellipses comme stimuli [Kosara *et al.* 2002b]. Les résultats de ces études ont mis en avant un phénomène d'attraction involontaire vers une ellipse nette parmi des ellipses floues. Le but de l'expérience qui suit est dans un premier temps de confirmer cet effet avec des stimuli plus complexes et plus naturels comme des mots écrits en toutes lettres. Cette étude compare aussi plusieurs niveaux de flou visuel afin d'en sélectionner un qui conduira à des performances similaires à celles obtenues pour une présentation auditive seule.

### 6.4.1 Protocole expérimental spécifique à l'expérience 1

L'expérience suivait un plan expérimental factoriel croisé inter-participants avec pour variables indépendantes : le niveau de flou (6 valeurs de $\boldsymbol{\sigma}$) et la congruence (**Neutre**, **Congruent**, **Incongruent**, **Dégradé**). Les variables dépendantes enregistrées incluaient le nombre d'erreurs et les RT.

La position de la cible était contrebalancée sur les 6 positions possibles. L'expérience totale était donc composée de 288 essais, c'est-à-dire 2 types de cible $\times$ 6 positions $\times$ 4 conditions de congruence $\times$ 6 niveaux de flou. Les essais étaient répartis en 12 blocs de 24 essais et les participants pouvaient prendre des pauses entre les essais pour diminuer leur fatigue. L'expérience complète durait environ 25 minutes.

La tâche utilisée était celle décrite en section 6.2.4. Les stimuli originaux étaient ceux décrits en section 6.3.1.3. Les versions floues de ces stimuli étaient obtenues en filtrant les stimuli originaux par un filtre gaussien de noyau 70 px et de rayon $\boldsymbol{\sigma}$ variant entre 3 et 18 par pas de 3. Plus $\boldsymbol{\sigma}$ était élevé et plus le stimulus obtenu était flou. Les stimuli étaient alignés comme indiqué dans la section 6.2.5.

Douze étudiants volontaires ont participé à cette expérience (moyenne d'âge 26 ans ; 2 femmes ; 10 droitiers). Tous étaient francophones de naissance. Tous avaient une vue normale ou corrigée au besoin et aucun n'a indiqué de problème visuel connu. Les participants étaient testés chacun leur tour dans une pièce isolée acoustiquement et sous les mêmes conditions d'éclairage et de position par rapport à l'écran.

### 6.4.2 Résultats

Dans un premier temps nous avons vérifié que les résultats ne dépendaient pas du type de cible (6 ou 10) à l'aide d'une analyse de variance avec mesures répétées (ou RM-ANOVA pour *Repeated Measures ANalysis Of VAriances*). Nous n'avons trouvé aucune différence significative entre les deux cibles concernant le nombre d'erreurs ($F(1,11)= 2.52$, $p = 0.14$), ni concernant les RT ($F(1,11)= 0.15$, $p = 0.70$). Pour la suite des analyses présentées ici nous avons donc traité

les données sans tenir compte du type de cible. Les résultats globaux de cette expérience sont présentés en figure 6.13.

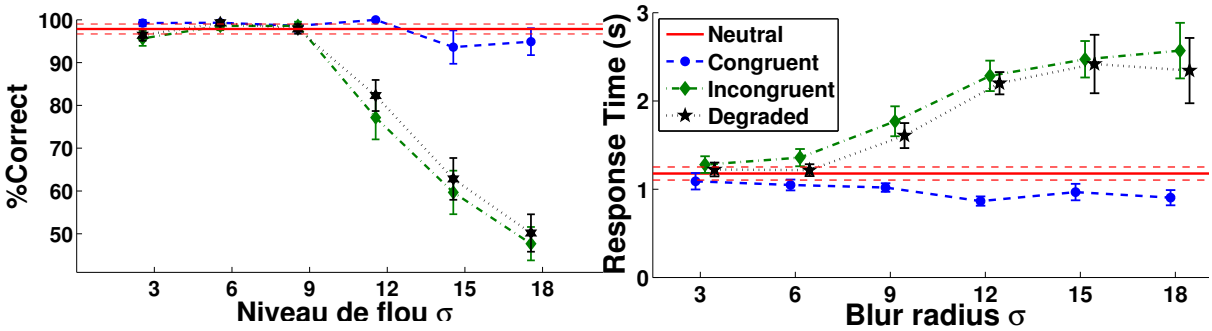



**Figure 6.13.** Taux de réponses correctes (gauche) et RT moyens (droite). Les barres d'erreurs représentent l'Erreur Type (*Standard Error of the Mean*). Le niveau de flou n'intervenant pas dans la condition Neutre, nous en avons moyenné les résultats.

#### 6.4.2.1 Taux de réponses correctes

Le taux de bonnes réponses, calculé comme le pourcentage de réponses correctes par participant sur l'ensemble des 6 essais d'une condition (une condition = un niveau de flou × une congruence), a été analysé par une RM-ANOVA à deux facteurs (niveau de flou et congruence). Les résultats montrent un effet significatif du niveau de flou ($F(5,55)= 54.11$, $p \ll 0.001$), un effet significatif de la congruence ($F(3,33)=75.71$, $p \ll 0.001$) et une interaction significative entre ces deux facteurs ($F(15,165)=23.68$, $p \ll 0.001$).

Un test *post hoc* de Duncan a confirmé que :

- l'effet de la congruence n'est pas significatif pour des niveaux de flous faibles ($\boldsymbol{\sigma}$ = 3, 6, 9) ;
- l'effet du niveau de flou n'apparaît que pour les niveaux de flous plus élevés ($\boldsymbol{\sigma}$ = 12,15,18) et pour les conditions **Incongruente** et **Dégradée**, c'est-à-dire quand la cible est floue : plus le niveau de flou augmente et plus le taux de bonnes réponses pour ces conditions diminue ($p \ll 0.001$). La différence entre les conditions où la cible est dégradée (**Incongruente**, **Dégradée**) et celles où la cible est nette (**Neutre**, **Congruente**) est alors significative ($p \ll 0.001$)
- le niveau de flou n'a pas d'effet significatif sur les deux conditions de congruence **Congruente** et **Neutre** : le pourcentage de bonnes réponses est constant et reste extrêmement élevé ;
- aucune différence significative ne peut être observée entre les conditions **Congruente** et **Neutre** quel que soit le niveau de flou ($p's > 0.25$) ;
- aucune différence significative ne peut être observée entre les conditions **Incongruente** et **Dégradée** ($p's > 0.15$) : la perte dans le taux d'identification est due au niveau flou de la cible et ne dépend pas de l'état, flou ou net, des distracteurs.

#### 6.4.2.2 Temps de réponse

Une RM-ANOVA sur les RT révèle un effet significatif du niveau de flou ($F(5,55)=8.22$, $p \ll 0.001$) et de la congruence ($F(3,33)=45.35$, $p \ll 0.001$) ainsi qu'une interaction significative entre ces deux facteurs ($F(15,165)=10.10$, $p \ll 0.001$). Un test de Duncan a par ailleurs confirmé les résultats suivants :

- la congruence n'influence pas le RT pour des niveaux de flous faibles ($\boldsymbol{\sigma} = 3, 6$) ;
- pour des niveaux plus élevés, le temps mis par les participants pour identifier une source floue (conditions **Incongruente** et **Dégradée**) est significativement plus long que pour identifier une source nette (**Neutre** et **Congruente**). En outre, plus le niveau de flou augmente et plus l'écart est grand et le temps mis pour identifier une source floue est important. L'identification d'une source floue est de plus en plus difficile.
- la différence entre les conditions **Congruente** et **Neutre** est quasi significative ($p = 0.065$), avec des RT plus courts pour la condition **Congruente**. Cette tendance se confirme quelque soit le niveau de flou.
- les RT ont tendance à être plus longs pour la condition **Incongruente** que pour la condition **Dégradée** ($p's > 0.19$).

#### 6.4.2.3 Résumé

Les résultats obtenus sur les RT sont cohérents avec ceux obtenus pour les taux de réponses correctes. On retiendra :

- une tendance vers un gain de temps pour la condition **Congruente** par rapport à la condition **Neutre** ;
- une tendance vers une perte de temps supplémentaire pour la condition **Incongruente** par rapport à la condition **Dégradée** ;
- cependant les deux points précédents n'étant pas significatifs, on remarque surtout que les résultats dépendent uniquement de l'état de la cible (floue ou nette) et non de celui des distracteurs ;
- plus le niveau de flou de la cible augmente et plus il est difficile de l'identifier : les taux d'identification sont de plus en plus faibles et les RT de plus en plus longs.

### 6.4.3 Discussion sur l'expérience visuel seul

#### 6.4.3.1 Dégradation due au flou visuel

Sans surprise, nos résultats montrent qu'une cible nette (dans les conditions **Neutre** et **Congruente**) conduit à une meilleure identification (en termes de RT et d'exactitude) qu'une cible floue (conditions **Incongruente** et **Dégradée**). Cela confirme que le flou visuel peut être vu comme une dégradation (Hypothèse 1).

#### 6.4.3.2 Attraction involontaire vers un objet net

La tendance observée pour les RT de la condition **Congruente** à être plus courts que ceux de la condition **Neutre** confirme l'intérêt de rendre flou les distracteurs pour faciliter la recherche d'une cible parmi plusieurs images fixes. Ce résultat est cohérent avec les résultats précédents obtenus sur des stimuli de laboratoires tels que des ellipses [Kosara *et al.* 2002b], cependant ces résultats étaient plus marqués que ceux que nous avons obtenus.

De même, bien qu'observée uniquement sous forme de tendance, la perte de temps supplémentaire ainsi que la baisse des taux d'identification entre la condition **Incongruente** et la condition **Dégradée** indiquent que l'attraction vers l'objet net est involontaire (Hypothèse 3) : bien que connaissant la tâche et cherchant à la réaliser, les participants ne peuvent ignorer le distracteur net [Theeuwes 2004]. Ils ont donc besoin de temps supplémentaire pour inhiber ce stimulus et procéder ensuite à une recherche séquentielle dans le sous-ensemble des éléments flous restants. Afin de confirmer que le regard est attiré vers la zone nette de l'image, nous aurions pu utiliser une analyse oculométrique pour mesurer la trajectoire du regard comme proposé dans [Olivier Le Meur 2012]. La zone la plus saillante est en effet la première zone observée. Les analyses oculométriques de précédentes études confirment d'ailleurs que le regard est effectivement attiré par une zone nette et repoussé par les zones floues [Grabowecky *et al.* 2012].

#### 6.4.3.3 Influence du niveau de flou visuel

Les résultats de [Kosara 2001, p. 68] montraient qu'un niveau de flou trop faible, typiquement inférieur à $\boldsymbol{\sigma} = 7$, n'introduisait pas d'attraction involontaire. L'auteur recommandait [p. 32] d'utiliser des niveaux de flous de 11 ou 15 px par défaut. Seuls ces trois niveaux de flou étaient alors utilisés. Or les résultats de notre expérience 1 montrent qu'un effet d'attraction involontaire n'est pas observable pour $\boldsymbol{\sigma} = 3$ ou 6 mais qu'un rayon de flou supérieur ($\boldsymbol{\sigma} = 9$, 12, 15 ou 18) produit un effet d'attraction involontaire influençant les RT. Ces valeurs sont donc cohérentes avec les résultats obtenus précédemment par Kosara.

#### 6.4.3.4 Différence de temps de réponse par rapport à l'état de l'art : rôle de la phase de décision

Dans l'expérience de Kosara et ses collègues [Kosara *et al.* 2002b], 63 stimuli étaient présentés en parallèle. Bien que nous n'ayons utilisé que 6 stimuli, les RT que nous avons mesurés sont nettement supérieurs à ceux qu'ils avaient obtenus (environ 650 ms dans l'expérience de Kosara). Premièrement, cela peut être dû au type de stimuli que nous utilisons. Composés de plusieurs lettres, nos stimuli ont également une composante sémantique et sont donc plus complexes et plus variés que les ellipses de Kosara. Cependant, dans l'expérience présentée par [Iordanescu *et al.* 2011] où la tâche était de chercher un objet visuel, sous forme d'une image ou d'un mot parmi 8 objets, les RT enregistrés étaient également de l'ordre de 650 ms. Nous supposons donc que la différence de temps de réponse entre ces expériences et la nôtre, est due à la tâche elle-même. Ils demandaient de retrouver une cible précise et connue (sans ambiguïté), tandis que nous demandions de retrouver la cible mais également de l'identifier parmi une des deux options de cible possibles (SIX ou DIX).

## 6.5 Expérience 2 : Effet du niveau de flou audio sur une recherche audio

Une seconde expérience a été menée pour évaluer l'effet d'attraction du flou audio et la possibilité de l'utiliser comme guide attentionel dans une tâche de recherche sonore. Nous espérions au moins le même niveau d'effet que dans l'expérience précédente sur du visuel seul ou dans celles décrites dans d'autres travaux.

### 6.5.1 Protocole expérimental spécifique à l'expérience 2

Afin de comparer les résultats de l'expérience 1 sur du visuel avec ceux de cette nouvelle expérience sur de l'audio seul nous avons gardé le même protocole expérimental dans les deux expériences en remplaçant les stimuli visuels par des stimuli audio. Le plan d'expérience suivait un design factoriel croisé inter-participants avec pour variables indépendantes : le niveau de flou (6 valeurs de $\boldsymbol{f_c}$) et la congruence (**Neutre**, **Congruent**, **Incongruent**, **Dégradé**).

La tâche utilisée est celle décrite en section 6.2.4. Les stimuli sont ceux décrits section 6.3.2. Des versions floues de ces stimuli audio sont créées en appliquant un filtre passe-bas de type Butterworth de $6^{\text{ième}}$ ordre. Six niveaux de flou sont simulés en appliquant six fréquences de coupure $\boldsymbol{f_c}$ différentes : 12500 Hz, 8500 Hz, 5000 Hz, 3500 Hz, 2500 Hz, and 1500 Hz. Le signal audio est de plus en plus flou lorsque la fréquence de coupure diminue. Les stimuli audio sont joués en boucle jusqu'à ce que le participant réponde et ils sont spatialisés comme expliqué dans la section 6.2.5.

Quinze étudiants et membres de l'Université Paris-Sud ont participé bénévolement à cette étude. Tous avaient une bonne audition et aucun n'a indiqué avoir de trouble auditif. Aucun des participants de cette expérience 2 "audio-seul" n'avait participé à la première expérience "visuel seul". Les participants étaient testés individuellement dans la même pièce que pour l'expérience précédente.

### 6.5.2 Résultats

Les résultats de cette expérience sont présentés en figure 6.14. Une RM-ANOVA à un facteur a révélé un effet significatif du type de cible ('6" ou "10", c'est-à-dire /*sis*/ ou /*dis*/) à la fois pour le taux de réponses correctes ($F(1,14)=14.16$, $p = 0.002$) et pour les RT ($F(1,14)=5.21$, $p = 0.038$). Nous avons donc séparé les analyses suivantes selon le type de cibles.

#### 6.5.2.1 Taux de réponses correctes

Les données concernant le %Correct diffèrent beaucoup en fonction de la cible. Une RM-ANOVA à 3 facteurs (cible, niveau de flou et congruence) a confirmé que les trois niveaux d'interactions (cible × niveau de flou, cible × congruence et cible × niveau de flou × congruence) avaient un effet significatif (F's>7.70, $p \ll 0.001$) : cette analyse montre que les effets des paramètres de niveau de flou et de congruence diffèrent selon la cible. Nous noterons d'ailleurs que le graphique obtenu pour la cible "6" (figure 6.14 en haut à gauche) se rapproche beaucoup

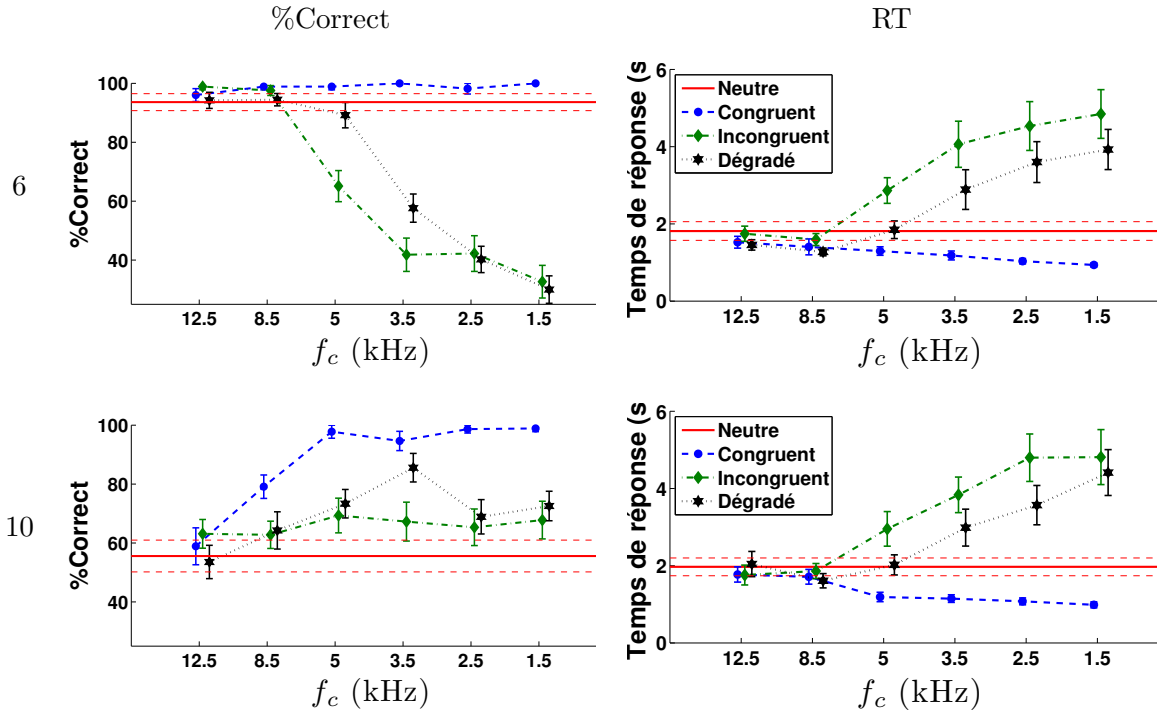



**Figure 6.14.** Taux de réponses correctes (gauche) et RT moyens (droite). Les barres d'erreurs représentent l'Erreur Type (*Standard Error of the Mean*). Le niveau de flou n'intervenant pas dans la condition neutre, nous en avons moyenné les résultats.

du graphique obtenu pour l'expérience 1 avec une présentation visuelle (figure 6.13, page 139), ce qui n'est pas le cas du graphique de résultat obtenu pour la cible "10".

La principale différence entre le "6" et le "10" apparaît pour la condition **Neutre**. Tandis que le "6" est correctement identifié ($\%Correct \sim 93.6$ %), le "10" semble moins facile à identifier ($\%Correct \sim 55.6$ %). Comme la tâche était de choisir entre le "6" et le "10", ces résultats semblent indiquer que les participants ont répondu "6" plus souvent que "10" dans cette condition **Neutre**.

Par ailleurs, le nombre de réponses correctes était plus élevé pour la condition **Congruente** que pour la condition **Neutre** confirmant notre **Hypothèse 2**. Cette amélioration d'identification n'était pas significative pour le nombre "6" ($p's \geq 0.1$), mais l'était pour le nombre "10" dès que le flou était suffisant ($p's \ll 0.0014$ dès que $f_c < 8500$ Hz), permettant de passer de seulement 58.9 % de taux d'identification en **Neutre** à 98.9 % en condition **Congruente**.

Cependant, bien que l'Hypothèse 2 soit confirmée pour le "10", nous observons que l'Hypothèse 1 ne l'est pas : de façon surprenante les conditions **Incongruente** et **Dégradée** conduisent à un meilleur taux d'identification que le **Neutre**. L'Hypothèse 1 reste cependant vérifiée pour le "6" où l'on observe une décroissance du taux d'identification en condition **Incongruente** et **Dégradée** lorsque le niveau de flou de la cible augmente.

De plus, pour les deux cibles on observe une différence significative entre les conditions

**Incongruente** et **Dégradée** avec le %Correct en **Incongruent** moins élevé pour $f_c = 3500\ Hz$ avec le "10" ($p \ll 0.001$) et pour $f_c = 5000$ et $3500\ Hz$ pour le 6 ($p \ll 0.001$). Une cible floue est donc encore moins facile à identifier si un des distracteurs était laissé net.

#### 6.5.2.2 Temps de réponse

Les données de RT ont été analysées par une RM-ANOVA avec les trois mêmes facteurs que pour l'analyse des réponses correctes. Les effets du niveau de flou ($F(5,70)=23.02$, $p \ll 0.001$), de la congruence ($F(3,42)=33.57$, $p \ll 0.001$) et de l'interaction niveau de flou × congruence ($F(15,210)=20.13$, $p \ll 0.001$) se sont révélés significatifs. Cependant, bien que nous ayons trouvé un effet de la cible sur les RT, nous n'avons pas trouvé d'interaction cible × niveau de flou ($F(5,70)=0.73$, $p = 0.60$), cible × congruence ($F(3,42)=0.93$, $p = 0.43$), ni cible × niveau de flou × congruence ($F(15,210)=0.45$, $p = 0.96$). Les RT pour le "10" semblent juste être légèrement plus longs que les RT pour le "6" et ce dans toutes les conditions de flou et de congruence.

L'effet de la congruence n'est pas observable pour des niveaux de flou faibles ($\boldsymbol{f_c}$=12500 ou 8500 Hz), $p's > 0.3$, mais pour des niveaux de flou plus élevé, un test *post hoc* de Duncan avec le niveau de flou et la congruence comme paramètre a révélé :

- des RT plus longs pour la condition **Incongruente** (pour tous les $\boldsymbol{f_c}$) et pour la condition **Dégradée** (pour $\boldsymbol{f_c} \leq 3500\ Hz$) par rapport à la condition **Neutre** : les participants ont besoin de plus de temps pour identifier la cible et prendre une décision concernant l'option de cible quand celle-ci est floue, donc le flou audio agit bien comme une dégradation (Hypothèse 1), $p's \ll 0.001$ ;
- des RT plus courts pour la condition **Congruente** comparée aux autres conditions, indiquant un effet de facilitation pour détecter et identifier la cible quand les distracteurs sont flous (Hypothèse 2), $p's < 0.01$. Cet effet de facilitation est renforcé quand le niveau de flou augmente : le gain en RT est encore plus grand, par exemple pour $\boldsymbol{f_c}$=1500 Hz, les RT sont quasiment divisés par deux et passent de $\simeq 1900\ ms$ pour le **Neutre** à seulement $\simeq 950\ ms$ pour la condition **Congruente** ;
- des RT plus longs pour la condition **Incongruente** que pour la condition **Dégradée** ($p's < 0.01$). Cette perte de temps supplémentaire ne peut être due à la dégradation de la cible mais peut être expliquée par un effet d'interférence dû au distracteur net. Il y a un effet de capture attentionnelle involontaire par ce distracteur et les participants ont besoin de plus de temps pour inhiber le traitement de ce stimulus et revenir au sous-ensemble constitué des autres sources (Hypothèse 3).

### 6.5.3 Discussion sur l'expérience audio seul

#### 6.5.3.1 Séparation de bandes fréquentielles et ségrégation

Il avait déjà été montré que dans le cas de plusieurs locuteurs simultanés, un filtrage fréquentiel différent pour chacune des voix améliore la ségrégation et augmente ainsi l'intelligibilité perçue. Par exemple, on pourra appliquer un filtrage passe-bas, un filtrage passe-bande et un

filtrage passe-haut pour séparer 3 locuteurs [Spieth *et al.* 1954]. Cela est en cohérence avec nos résultats puisque nous observons un meilleur taux d'identification pour une cible qui contient plus de fréquences aiguës que les distracteurs. Ainsi la netteté d'une source peut en faciliter le suivi (attention volontaire).

#### 6.5.3.2 Attraction involontaire vers la source nette

Le démasquage, c'est-à-dire la capacité à augmenter l'intelligibilité et la ségrégation d'une source initialement masquée par plusieurs sons masquants, est une première étape pour la conception d'une interface de présentation de plusieurs sources sonores concurrentes. Notre étude va cependant plus loin : il ne s'agit pas seulement ici de faciliter la ségrégation mais bien d'orienter l'utilisateur vers un document sélectionné parmi plusieurs documents. Nous avons également trouvé un effet d'attraction involontaire sur l'objet net qu'il soit pertinent pour la tâche (la cible en condition **Congruente**) ou non (un distracteur en condition **Incongruente**) : une source audio nette (contenant des hautes fréquences) attire l'attention quand elle est située en concurrence avec d'autres sources sonores floues.

#### 6.5.3.3 Utilisation du flou audio en IHM

On pourrait se servir avantageusement de cette technique dans les interfaces homme-machine et les systèmes de communication pour plusieurs locuteurs (armée [Brungart et Simpson 2005], visioconférence...). Du moment que l'auditeur sait sur quel flux audio porter son attention, il lui suffit alors d'appliquer un filtre aux autres flux pour mettre le premier en valeur. Cette technique de guide attentionnel pourrait aussi être utilisée dans le design d'interfaces audio dédiées aux mal-voyants et non-voyants car il facilite la recherche d'un objet sonore sans en impliquer la vision d'une représentation textuelle ou iconique.

Cependant, l'intelligibilité d'une source floue est considérablement diminuée comme le montre les performances des conditions **Dégradée** et **Incongruente**. Si le flou audio appliqué aux distracteurs facilite la ségrégation et le suivi d'une source nette, il ne peut cependant pas être utilisé dans des tâches d'attention divisée où il est important de pouvoir suivre et comprendre deux flux audio simultanés.

Cette technique semble surtout utilisable dans le cas où l'utilisateur peut contrôler quelle information mettre en valeur. Il peut ainsi modifier son point de focus de manière interactive en augmentant la netteté d'un autre flux sonore qu'il souhaite explorer. Nous pouvons imaginer un système de lentille audio qui reprend le principe de la ***SoundTorch*** [Heise *et al.* 2008] ou de la lentille grossissante présentée au chapitre 4 mais qui, au lieu d'augmenter le niveau sonore de la source sous le curseur, appliquerait un filtrage (ou plus généralement un flou audio) sur les sources en dehors de la lentille. On peut aussi imaginer un système où les deux paramètres taille et netteté sont combinés pour faire d'autant plus ressortir l'information sous le curseur.

#### 6.5.3.4 Asymétrie 6 vs. 10

Le %Correct (pourcentage de réponses correctes) était généralement plus bas pour le “10” que pour le “6”. Du fait que l'expérience était basée sur un choix forcé à deux options seulement, la différence dans le %Correct indique que les participants ont répondu “6” plus souvent

que “10”, en particulier quand la cible était nette ou seulement légèrement floue. L'asymétrie dans les réponses s'inversait cependant pour des niveaux de flou plus élevés et les participants répondaient alors plus souvent “10” que “6”. Dans un premier temps nous avons supposé que le filtrage des hautes fréquences avaient plus d'influence sur le son /*s*/ que sur le son /*d*/ puisque le son /*s*/ contient naturellement plus de hautes fréquences. Cependant nous pouvons supposer également que le masquage informationnel explique aussi, au moins partiellement, l'asymétrie entre les deux types de cible puisque certains des distracteurs commencent par le même son que la cible mais plus souvent par le son /*s*/ (/*sis*/ “6”, /*sɛ̃k*/ “5”, and /*sɛt*/ “7”) que par le son /*d*/ (/*dis*/ “10” and /*d*/ “2”). Par ailleurs, des effets de fusion entre stimuli peuvent avoir lieu, par exemple le son /*d*/ du “2” et le son /ˈɥit/ du “8” joués simultanément avec un faible écart spatial ressemblent au son /*dis*/ (“10”) même si la cible “6” est présente. L'asynchronie entre les différents stimuli, notamment aux niveaux des attaques, devrait réduire le problème de fusion 3.3.1 et [Bregman 1990]. De plus, comme l'enveloppe temporelle et, à plus haut niveau, les aspects sémantiques sont aussi des paramètres permettant la ségrégation des sources, le problème de fusion devrait être diminué avec des documents audio plus longs et au contenu sémantique plus riche.

#### 6.5.3.5 Influence du filtrage sur la spatialisation et la séparation de sources

D'après Carlile et Schönstein [Carlile et Schonstein 2006], les hautes fréquences favorisent la séparation spatiale de sources sonores et augmentent ainsi l'intelligibilité de la parole dans des conditions de locuteurs multiples. Bien que les auteurs ne relèvent une contribution de ces HF que pour l'élévation et non pour la séparation en azimut, le filtrage passe-bas que nous appliquons pour flouter les distracteurs pourrait réduire les capacités de l'utilisateur à localiser et donc à distinguer les six stimuli concurrents. Cet effet pourrait en partie expliquer les différences entre condition **Dégradée** et condition **Neutre** puisque l'on floute l'ensemble des stimuli dans la condition **Dégradée**, mais ne peut expliquer l'effet de facilitation observé pour la condition **Congruente** ni la perte d'efficacité (effet de capture attentionnelle) pour la condition **Incongruente** puisque le même nombre de stimuli est flou dans ces deux conditions, conduisant donc dans les deux cas à la même ségrégation spatiale.

## 6.6 Calibration multimodale : Comparaison et sélection d'un niveau de flou audio et visuel

D'après le modèle de maximum de vraisemblance et le principe d'efficacité inverse (chapitre 3, section 3.4.2.1b), lorsque les deux modalités ne donnent pas de réponses du même ordre de grandeur en termes d'intensité perçue ou de temps de réponse, la modalité la plus fiable va dominer et l'autre modalité ne sera quasiment pas prise en compte. Pour éviter d'introduire un biais vers l'une ou l'autre des modalités, nous avons donc cherché à calibrer l'effet du flou audio et du flou visuel. Pour cela, nous avons comparé les résultats obtenus entre la première expérience, visuel seul, et la seconde expérience, audio seul, de façon à trouver un niveau de flou visuel et un niveau de flou audio qui conduisent à des performances comparables entre ces deux modalités.

### 6.6.1 Comparaison des résultats en visuel et en audio

Dans les expériences 1 et 2, correspondant aux modalités visuelles et auditives, les RT ne dépendaient pas de la cible et suivaient le même profil pour le « 6 » et le « 10 » et pour les deux modalités. En revanche, contrairement aux résultats visuels de l'expérience 1, les résultats audio de l'expérience 2 concernant le %Correct présentaient des différences significatives entre les deux options de cible. Ainsi seul le « 6 » conduisait à des résultats similaires en audio et en visuel. Dans la suite de l'étude, nous devrons donc différencier les deux cibles lors des analyses.

### 6.6.2 Sélection d'un niveau de flou

La prochaine étape de cette étude combine et compare les flous audio et visuels afin d'évaluer l'éventuel gain obtenu dans une tâche de recherche bimodale éventuellement facilitée par un trait caractéristique lui-même bimodal. Pour procéder à cette étape, il fallait d'abord trouver un niveau de flou audio et un niveau de flou visuel conduisant :

1. chacun à un effet de facilitation unimodal où une cible nette est retrouvée plus vite au sein d'un groupe de distracteurs flous qu'au sein d'un groupe de distracteurs nets. En visuel, cette condition était obtenue pour les niveaux de flou $\boldsymbol{\sigma}$= 9, 12, 15, et 18. En audio, elle était vérifiée pour les niveaux de flou $\boldsymbol{f_c}$= 5000, 3500, 2500, et 1500 Hz ;
2. à des performances comparables en termes de taux %Correct et/ou de RT. Bien que les tendances soient similaires pour chaque modalité prise individuellement, les profils de courbes ne correspondent pas à la fois pour le %Correct et pour les RT. Seuls les RT conduisent à des profils similaires entre les deux modalités. On notera tout de même que les différences de temps entre les conditions **Congruente** et **Neutre** (gain pour la condition **Congruente**) et entre les conditions **Incongruente** et **Dégradée** (perte pour la condition **Incongruente**) étaient toutes significatives en audio, tandis que seule la différence entre les conditions **Congruente** et **Neutre** l'était en visuel.

Ainsi, à la suite de cette phase de calibration, il nous a été impossible de trouver une calibration parfaite qui permette d'aligner les performances visuelles et auditives à la fois sur les RT et sur le %Correct. Comme nous voulions évaluer l'aspect involontaire de l'attraction provoquée par le net et que les RT semblent plus pertinents pour mesurer ces processus perceptifs involontaires [Prinzmetal *et al.* 2005], nous avons préféré sélectionner les deux niveaux de flous correspondant à des RT similaires entre les deux modalités.

De plus, les différences entre les conditions de congruence n'étaient pas tout à fait équivalentes entre les deux modalités. Nous avons arbitrairement sélectionné les niveaux de flous par rapport aux RT mesurés dans la condition **Dégradée** pour les deux modalités. Les RT ne sont alors pas exactement égaux mais ils sont comparables (du même ordre de grandeur) entre les deux modalités et ce pour toutes les conditions de congruence. Nous avons ainsi pris pour niveau de flou audio $\boldsymbol{f_c}$ = 5000 Hz et pour niveau de flou visuel $\boldsymbol{\sigma}$ = 12. Ces deux niveaux de flous conduisaient à des RT d'environ 2 s dans chaque modalité. Ce niveau de flou visuel est cohérent avec les études faites par [Kosara 2001, p.32] puisqu'il recommande d'utiliser un niveau de flou entre 11 et 15 pixels.

## 6.7 Expérience 3 : Comparaison et combinaison des flous audio et visuel pour le guidage attentionnel multimodal

Les deux expériences précédentes ont confirmé que la netteté visuelle et la netteté auditive sont deux traits caractéristiques unimodaux qui facilitent la recherche visuelle d'une part, et auditive d'autre part. De plus, d'après le chapitre 3, une présentation bimodale améliore les performances. Toutefois pour ce qui est d'une recherche audiovisuelle, nous ne savons pas comment les traits caractéristiques de chaque modalité interagiront. L'expérience qui suit a donc été développée pour tester nos hypothèses sur une présentation bimodale. En plus des questions déjà traitées dans les deux expériences précédentes sur l'effet de dégradation et l'effet d'attraction involontaire, cette expérience a pour but de répondre aux questions suivantes :

1. Quelle est la contribution de chaque modalité dans une recherche bimodale ? (Hypothèse 4 sur l'*effet de cibles redondantes*)
2. Quel est l'effet de la combinaison de deux traits caractéristiques issus de modalités différentes ? (Hypothèse 5 sur l'*effet de redondance de traits caractéristiques*)

Contrairement aux deux expériences précédentes qui comparaient différents niveaux de flous, cette expérience compare différentes combinaisons des modalités audio et visuelles. Le protocole expérimental est donc différent.

### 6.7.1 Protocole expérimental spécifique à l'expérience 3

#### 6.7.1.1 Participants

Douze personnes (âge moyen 26 ans ; 4 femmes, 8 hommes ; tous droitiers ; tous francophones) ont participé à cette expérience. Aucune d'entre elles n'avait participé aux expériences 1 et 2. Toutes avaient une vue normale ou corrigée sans historique de problèmes de vision. Les participants passaient auparavant un test audiométrique attestant qu'ils n'avaient pas de problème auditif (niveau audiométrique normal ($< 20$ dB HL) par bande d'octaves entre 250 et 8000 Hz). Les participants étaient testés individuellement dans la même pièce et sous les mêmes conditions d'éclairage que pour les expériences 1 et 2.

#### 6.7.1.2 Stimuli

Les stimuli utilisés dans cette expérience étaient sélectionnés parmi les stimuli visuels de l'expérience 1 et les stimuli auditifs de l'expérience 2. L'ensemble de stimuli utilisés ici était donc constitué des stimuli originaux nets (4 exemplaires $\times$ 10 nombres prononcés pour l'audio + 4 exemplaires $\times$ 10 nombres écrits pour le visuel) et des stimuli flous obtenus pour un seul niveau de flou (un niveau visuel et un niveau audio). Ce niveau de flou a été sélectionné en comparant les résultats des expériences 1 et 2 (voir section 6.6) pour que les deux niveaux de flous conduisent à des RT comparables entre l'audio et le visuel pour une condition **Dégradée**. Les versions floues sont donc obtenues en appliquant un filtre Gaussien de noyau 70 px et de rayon $\boldsymbol{\sigma} = 12$ px pour le visuel, et un filtre passe-bas de fréquence de coupure $\boldsymbol{f_c} = 5000$ Hz pour l'audio.

**Table 6.1.** Représentation symbolique d'un exemple de chaque condition du design factoriel entre condition de congruence et la modalité employée. La cible (ici 6) est soulignée, les objets nets sont représentés en gras, les objets flous en italique. Les composantes visuelles et sonores sont séparées.

| | | Congruence | | | |
|---|---|---|---|---|---|
| Modalité | Composante | Neutre | Congruent | Incongruent | Dégradé |
| Audio-seul ($A$) | A | **1 2 3 4 5 6** | *1 2 3 4 5 6* | *1* **2** *3 4 5 6* | *1 2 3 4 5 6* |
| | V | | | | |
| Visuel-seul ($V$) | A | | | | |
| | V | **1 2 3 4 5 6** | *1 2 3 4 5 6* | *1* **2** *3 4 5 6* | *1 2 3 4 5 6* |
| Audiovisuel | A | **1 2 3 4 5 6** | *1 2 3 4 5 6* | *1* **2** *3 4 5 6* | *1 2 3 4 5 6* |
| flou bimodal ($AV$) | V | **1 2 3 4 5 6** | *1 2 3 4 5 6* | *1* **2** *3 4 5 6* | *1 2 3 4 5 6* |
| Audiovisuel | A | **1 2 3 4 5 6** | *1 2 3 4 5 6* | *1* **2** *3 4 5 6* | *1 2 3 4 5 6* |
| flou audio ($AV^{neut}$) | V | **1 2 3 4 5 6** | **1 2 3 4 5 6** | **1 2 3 4 5 6** | **1 2 3 4 5 6** |
| Audiovisuel | A | **1 2 3 4 5 6** | **1 2 3 4 5 6** | **1 2 3 4 5 6** | **1 2 3 4 5 6** |
| flou visuel ($A^{neut}V$) | V | **1 2 3 4 5 6** | *1 2 3 4 5 6* | *1* **2** *3 4 5 6* | *1 2 3 4 5 6* |

#### 6.7.1.3 Plan d'expérience et procédure

Nous avons utilisé un plan d'expérience factoriel croisé $4 \times 5$ intra-participants avec pour variables indépendantes : la congruence (**Neutre**, **Congruente**, **Incongruente**, **Dégradée**) et la combinaison de modalité (**A**, **V**, **AV**, $\mathbf{AV^{neut}}$, $\mathbf{A^{neut}V}$). Les conditions de congruence sont les mêmes que celles utilisées dans les expériences 1 et 2. Concernant les combinaisons de modalité, nous avons testé deux conditions unimodales et trois conditions bimodales. En effet, nous voulions évaluer le rôle de chacune des deux modalités dans une tâche de recherche bimodale et vérifier nos hypothèses concernant la redondance de cible (Hypothèse 4). Nous avons donc comparé la modalité audio-seul (**A**) et la modalité visuel-seul (**V**) à une combinaison bimodale où un trait caractéristique (net *vs* flou) était appliqué aux deux modalités de façon cohérente **AV**. De plus, pour tester l'effet de redondance de trait caractéristique (*Redundant Cueing Effect*, Hypothèse 5) et évaluer la contribution de chaque trait caractéristique unimodal (flou audio *vs* flou visuel) dans la combinaison redondante (flou audiovisuel), nous avons également testé deux autres combinaisons bimodales, une où le traitement n'est appliqué qu'à l'audio et tous les stimuli visuels restent nets (comme si le visuel était neutre) $\mathbf{AV^{neut}}$ et une où le traitement n'est appliqué qu'au visuel et tous les stimuli sonores restent nets $\mathbf{A^{neut}V}$. La Table 6.1 présente une schématisation des différentes conditions expérimentales. La vidéo 6.1 6.1 expose le rendu audiovisuel présenté aux participants suivants les différentes conditions.

La tâche était la même que dans les deux premières expériences : les participants devaient déterminer laquelle des deux options de cible ("6" ou "10") était présente parmi les six stimuli concurrents.

Pour les essais bimodaux **AV** , $\mathbf{A^{neut}V}$ et $\mathbf{AV^{neut}}$, les stimuli étaient présentés aux participants avec leur deux composantes, auditive et visuelle, simultanément. Les participants étaient

**Table 6.2.** Taux de réponses correctes par participant (%Correct), moyenne (erreur standard) des participants selon les différentes conditions.

| | | Modalité | | | | |
|---|---|---|---|---|---|---|
| Congruence | Cible | $A$ | $V$ | $AV$ | $AV^{neut}$ | $A^{neut}V$ |
| Neutre | 6 | 93.8 (2.2) | 97.5 (1.2) | 98.4 (1.1) | 99.3 (0.7) | 96.4 (1.5) |
| | 10 | 45.2 (4.7) | 98.1 (1.0) | 95.1 (1.9) | 95.0 (1.6) | 94.2 (2.2) |
| Congruent | 6 | 97.4 (1.2) | 96.7 (1.7) | 98.2 (1.3) | 96.8 (1.7) | 99.3 (0.7) |
| | 10 | 96.4 (1.5) | 98.5 (1.1) | 98.2 (1.3) | 97.1 (1.7) | 94.1 (2.4) |
| Incongruent | 6 | 76.8 (3.8) | 91.6 (2.0) | 81.3 (3.9) | 96.9 (1.5) | 92.5 (2.7) |
| | 10 | 66.3 (4.2) | 93.8 (2.2) | 86.4 (3.6) | 97.2 (1.3) | 77.5 (4.2) |
| Dégradé | 6 | 87.4 (3.7) | 87.5 (3.5) | 90.1 (3.5) | 97.4 (1.5) | 86.7 (3.6) |
| | 10 | 77.6 (3.6) | 94.5 (1.9) | 88.6 (2.9) | 97.4 (1.3) | 79.2 (3.6) |

alors libres de se fier à leur audition, à leur vision ou à une combinaison de leurs deux sens. La position visuelle des stimuli était la même que dans l'expérience 1 et la spatialisation des stimuli sonores la même que dans l'expérience 2 comme expliqué section 6.2.5. Ainsi les stimuli visuels étaient statiques et les stimuli auditifs étaient lus en boucle jusqu'à la réponse du participant.

Au total chaque participant effectuait 2 cibles × 6 positions × 4 congruences × 5 modalités soit 240 essais. Les essais étaient groupés en 10 blocs de 24 essais. Les participants pouvaient prendre des pauses entre les blocs pour limiter les effets de la fatigue. Afin de diminuer les effets de switch attentionnel d'une modalité à l'autre [Lukas *et al.* 2010], chaque bloc ne faisait intervenir qu'une seule modalité : audio-seul, visuel-seul ou audiovisuel. Toutefois, pour que, dans les essais bimodaux avec un trait caractéristique unimodal ($\mathbf{A^{neut}V}$ et $\mathbf{AV^{neut}}$), les participants ne sachent pas à l'avance sur quelle modalité se fier, nous avons mélangé les trois conditions bimodales au sein des blocs audiovisuels. Les participants étaient informés de la modalité employée dans chaque bloc. Au final ils étaient testés sur 2 blocs audio-seul ($\mathbf{A}$), 2 blocs visuels seuls ($\mathbf{V}$) et 6 blocs multimodaux ($\mathbf{AV}$ , $\mathbf{A^{neut}V}$ et $\mathbf{AV^{neut}}$). L'ordre de présentation des modalités était contrebalancé par un carré latin à 3 rangs. De façon à augmenter le nombre d'observations, l'expérience était répétée deux fois pour chaque participant. Avant l'expérience les participants s'exerçaient sur un bloc d'apprentissage avec 12 essais de chaque modalité présentés dans l'ordre assigné par le carré latin. L'expérience totale durait environ 40 minutes.

### 6.7.2 Résultats

Comme pour les expériences, nous avons analysé deux variables dépendantes sur les performances des participants : les taux de réponses correctes (%Correct) et les temps de réponse (RT).

#### 6.7.2.1 Taux de réponses correctes

Les %Correct moyens sont reportés en Table 6.2. Une RM-ANOVA a été réalisée sur les trois facteurs type de cible, congruence et combinaison de modalité. D'après l'analyse, la congruence

(F(3,33)=40.55, $p \ll 0.001$) et la modalité (F(4,44)=39.88, $p = 0.001$) affectent significativement les résultats. Leur interaction est également significative (F(12,132)=11.65, $p < 0.001$). De plus, l'analyse confirme une forte différence entre les cibles "6" et "10" (F(1,11)=18.98, $p = 0.001$) et des interactions significatives type de cible × congruence (F(3,33)=7.00, $p < 0.001$), type de cible × modalité (F(4,44)=11.14, $p \ll 0.001$) et type de cible × congruence × modalité (F(12,132)=5.07, $p \ll 0.001$). Nous avons donc analysé séparément les résultats pour les deux cibles. Un test *post hoc* de Duncan a été effectué pour distinguer les effets de chaque facteur.

Pour la condition $\mathbf{AV^{neut}}$ avec le flou appliqué seulement à l'audio, aucune différence ne peut être observée ni entre les cibles, ni entre les différentes conditions de congruence ($p's > 0.53$), toutes les conditions menant alors à des %Correct élevés. De même, les deux cibles ne diffèrent pas l'une de l'autre pour les autres conditions où la vision intervient (**V**, **AV**, $\mathbf{A^{neut}V}$). Le %Correct est alors très élevé dans les conditions où la cible est nette (**Neutre** et **Congruente**) et plus faible lorsque la cible est floue (**Incongruente** et **Dégradée**). Cette différence est significative pour **AV**, $\mathbf{A^{neut}V}$ et **V-Dégradé** ($p < 0.02$) et quasi significative pour **V-Incongruent** ($p = 0.09$). Par ailleurs aucune différence significative ne peut être observée entre les cas **Incongruent** et **Dégradé** ($p's > 0.11$), ni entre les cas **Neutre** et **Congruent** ($p's > 0.7$). Ainsi le taux d'identification ne semble dépendre ici que de l'état flou ou net de la cible. Ces résultats confirment que le flou agit comme une dégradation (Hypothèse 1). Le %Correct, en présence de la composante visuelle, n'est pas affecté par un phénomène d'attraction involontaire.

Les deux cibles "6" et "10" donnent surtout des résultats différents pour l'audio (**A**). En particulier la condition **Neutre** conduit à des %Correct très faibles pour le "10" mais pas pour le "6". Cela signifie que les participants ont répondu plus souvent "6" que "10", même lorsque la cible était en fait un "6". En conséquence les conditions **Incongruente** et **Dégradée** conduisent à des taux d'identification significativement supérieurs ($p's \ll 0.001$) bien que la cible soit floue. Cette asymétrie n'apparaît pas dans les autres conditions de modalités et est donc intrinsèque à la modalité auditive. On remarque cependant que d'une part la congruence compense la perte d'identification pour le "10" (on passe de 45 % en **Neutre** à 96 % de bonnes réponses en **Congruent**). D'autre part, pour les deux cibles le %Correct de la condition **Incongruente** est inférieur à celui de la condition **Dégradée**. Enfin, cette baisse du taux d'identification n'apparaît pas dans la condition $\mathbf{AV^{neut}}$ où une composante visuelle nette est ajoutée. Dans ce cas où l'audio est dégradé et conduirait à de mauvaises performances individuellement, les participants se sont donc fiés au visuel.

#### 6.7.2.2 Temps de réponse

Nous avons analysé les RT en considérant l'ensemble des essais (bonnes et mauvaises réponses) afin de ne pas supprimer plus d'essais dans une condition que dans les autres (le "10" conduisant à 55 % d'erreur en audio **Neutre**). Nous avons cependant vérifié auparavant que les mauvaises réponses conduisaient à des RT plus longs pour toutes les conditions, ne favorisant pas une condition plus qu'une autre.

L'analyse est effectuée avec une RM-ANOVA et les mêmes facteurs que pour l'ana-

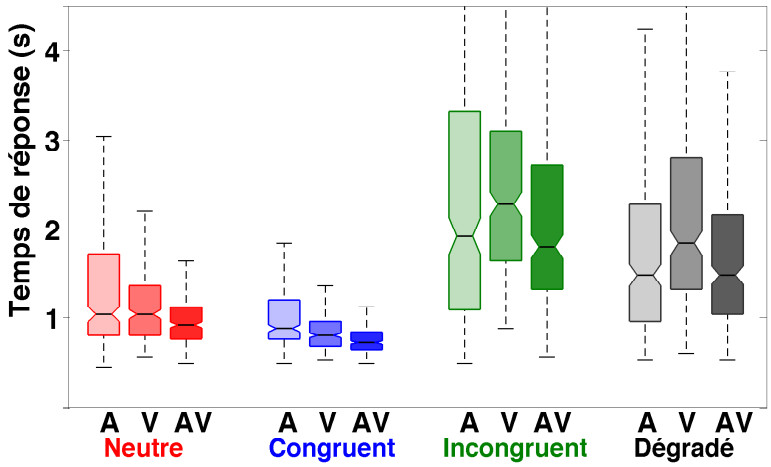



**Figure 6.15.** RT des conditions unimodales et de la condition bimodale cohérente. La représentation se fait sous forme de boîte à moustaches : la barre centrale est la médiane, les boîtes indiquent les 25$^{\text{ème}}$ et 75$^{\text{ème}}$ quantiles, les moustaches indiquent les données jusqu'à la valeur médiane±1.5 écart-type.

lyse des %Correct. Les effets de la modalité (F(4,44)=18.17, $p \ll 0.001$), de la congruence (F(3,33)=107.2, $p \ll 0.001$) et de l'interaction modalité × congruence (F(12,132)=11.73, $p \ll 0.001$) se sont tous révélés significatifs. En revanche, et contrairement au %Correct, le type de cible n'influence pas significativement les RT directement (F(1,11)=0.12, $p = 0.73$), et peu par interaction (cible × modalité F(4,44)=0.99, $p = 0.42$, cible × congruence F(3,33)=1.49, $p = 0.23$, cible × congruence × modalité F(12,132)=1.97, $p = 0.03$). Nous les présentons donc conjointement sans séparer le type de cible. Les analyses plus approfondies suivantes relèvent d'un test *post hoc* de Duncan.

#### a) Effet de la congruence et de la redondance de cible

La figure 6.15 présente les RT des présentations unimodales (**A** et **V**) et bimodale où le flou est appliqué aux deux modalités (**AV**). Elle permet donc d'évaluer l'effet de le redondance de cible. On observe :

**L'identification d'une cible floue est plus longue**. Les conditions **Incongruente** et **Dégradée** conduisent à des RT plus longs que ceux obtenus dans une condition **Neutre**, $p's < 0.001$. Il faut donc plus de temps pour retrouver et identifier une cible floue qu'une cible nette ce qui est cohérent avec les taux de réponses correctes et confirme notre Hypothèse 1.

**Un indice congruent avec la cible conduit à des RT plus courts**. Ce gain en temps par rapport à la condition **Neutre** s'observe, de manière significative ou quasi significative, pour toutes les modalités (**A**, **V**, **AV**). Ainsi les moyennes des RT passent, dans le sens **Neutre** à **Congruent**, de 1.71 s à 1.24 s pour **A** ($p \ll 0.001$), de 1.13 s à 0.94 s pour **V** ($p = 0.10$), et de 0.94 s à 0.77 s pour **AV** ($p = 0.08$) ce qui représente un gain de plus de 16 %. De plus on remarque une nette diminution des variances en condition **Congruente**, comme l'indique

la longueur des moustaches sur le graphique 6.15, ce qui confirme qu'il y a moins de risque d'obtenir des RT longs dans cette condition que dans les autres. Ces résultats sont cohérents avec notre Hypothèse 2.

**Un objet net attire involontairement l'attention parmi des objets flous**, qu'il soit pertinent ou non pour la réalisation de la tâche. Ainsi, la condition **Incongruente** conduit à des RT encore plus longs que ceux obtenus en condition **Dégradée**, $p's < 0.015$. Ce temps supplémentaire ne peut être dû uniquement à la difficulté d'identification de la cible floue. Il indique donc que les participants ont eu besoin de plus de temps lorsqu'un autre distracteur était net et donc saillant. Ce résultat est cohérent avec l'Hypothèse 3 selon laquelle le temps supplémentaire est en fait utilisé à inhiber le distracteur par lequel le participant est involontairement attiré.

**Une combinaison bimodale cohérente**[4] **facilite et raccourcit les temps de recherche**. Les RT des conditions **Neutre**, **Congruente** et **Incongruente** sont donc cohérents avec l'Hypothèse 4. En effet, on observe sur la figure 6.15 que la tâche est réalisée plus rapidement en présence des deux modalités et des deux indices **AV** que lors d'une présentation individuelle de chaque modalité visuelle **V** ou auditive **A**. Il y a donc effet de redondance de cible. La différence est significative pour toutes les conditions audio ($p's < 0.01$) et pour le visuel **Incongruent**. Les différences entre le **Neutre-V** et **Neutre-AV** ainsi qu'entre **Congruent-V** et **Congruent-AV** ne sont pas significatives ($p's \geq 0.24$). Toutefois dans tous les cas, on observe que les variances sont plus petites pour une condition bimodale **AV** que pour les conditions unimodales **A** et **V**. Le risque d'obtenir un temps de recherche long est donc limité dans le cas bimodal. Dans la condition **Dégradée**, les RT bimodaux ne sont cependant pas plus rapides que chacun des deux RT unimodaux : ils sont plus courts que les RT visuels ($p = 0.003$) mais équivalents aux RT audio ($p = 0.88$) qui correspondent alors aux RT les plus courts. Ainsi, la tâche de recherche bimodale est équivalente à la tâche de recherche la plus rapide, voire encore plus rapide.

#### b) Effet de la redondance intermodale de traits caractéristiques

La figure 6.16 permet d'évaluer la contribution de chaque trait caractéristique unimodal (flou auditif $\mathbf{AV^{neut}}$ et flou visuel $\mathbf{A^{neut}V}$) dans la combinaison bimodale (flou audiovisuel **AV**).

**Le flou audio seul ne suffit pas pour faciliter une recherche bimodale**. Dans la condition $\mathbf{AV^{neut}}$, on observe une absence de différence entre les conditions de congruence $p's > 0.61$. L'effet du flou audio, positif ou négatif selon que le flou est appliqué à la cible ou à un distracteur, n'apparaît plus en présence d'une présentation visuelle neutre (et donc fiable) alors qu'il apparaissait en l'absence de présentation visuelle (**A**).

**La redondance des traits caractéristiques audio et visuels n'apporte aucun gain.** Nous avons comparé les conditions bimodales avec le trait caractéristique présent dans les deux modalités **AV** ou appliqué au visuel seulement $\mathbf{A^{neut}V}$. Cela nous permet d'évaluer l'apport du flou audio en le distinguant de l'apport de la modalité auditive. Les conditions **Incongruente**

4. Par cohérence, on désigne ici le fait d'appliquer le trait caractéristique aux deux modalités de façon cohérente.

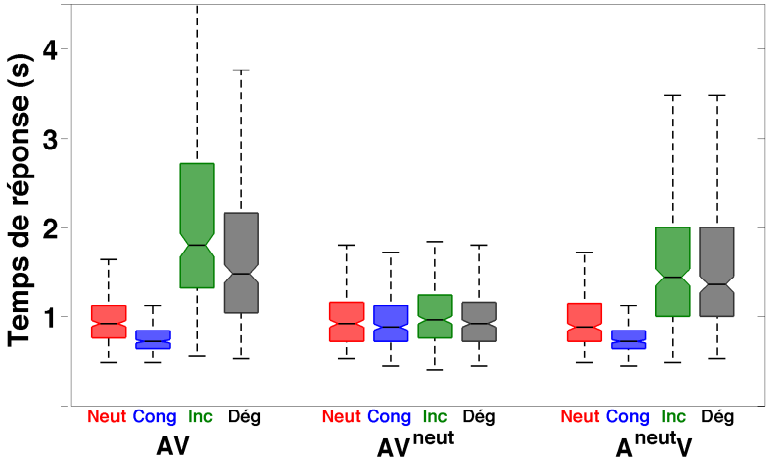



**Figure 6.16.** RT des conditions bimodales avec un trait caractéristique bimodal ou unimodal.

et **Dégradée** en $\mathbf{A^{neut}V}$ conduisent à des RT similaires, $p = 0.98$. Au contraire, ces conditions donnent des RT significativement différents pour **AV**, $p = 0.002$, avec la condition **Incongruente** plus lente. Ainsi l'indice audio dans **AV** renforce l'effet d'attraction du distracteur saillant par rapport au cas où il n'y pas d'indice audio ($\mathbf{A^{neut}V}$). Cependant, en terme de gain, il n'y pas de réel apport de l'indice audio dans la condition **Congruente** et les RT de **AV** et $\mathbf{A^{neut}V}$ ne diffèrent pas significativement $p = 0.83$. L'Hypothèse 5 n'est donc pas vérifiée pour ce type d'indice.

### 6.7.3 Discussion partielle

#### 6.7.3.1 Renforcement de la saillance d'un objet net : utilisation d'un trait caractéristique pour la recherche

Nous avions déjà vérifié, dans les expériences 1 et 2, que les flous audio et visuel, correspondant tous les deux à un filtrage des hautes fréquences, permettent de rendre la cible plus saillante quand le flou est appliqué aux distracteurs. Les résultats de l'expérience 3 confirment ces résultats et révèlent qu'un flou audiovisuel, obtenu par combinaison des flous unimodaux, met aussi la cible nette en valeur. Dans toutes les combinaisons de modalité, la netteté peut être utilisée comme trait caractéristique pour attirer l'attention vers la cible. De plus, pour toutes les modalités, la condition incongruente s'est révélée la plus lente : l'effet d'attraction ne peut donc pas être uniquement dû à la tâche (chercher la cible) mais est un processus involontaire. Cela est particulièrement intéressant pour la conception d'IHM car cela signifie que ce processus d'attention sélective demande moins d'effort à l'utilisateur si les distracteurs sont flous.

#### 6.7.3.2 Effet de redondance de cible : avantage de la bimodalité

De façon cohérente avec les études préalables en multimodalité et multisensorialité, les conditions bimodales conduisent à de meilleures performances en exactitude et en temps de réponse, ou au moins équivalentes à la meilleure des deux conditions unimodales. Plus précisément, le

gain en temps observé pour les conditions neutre et congruente en audiovisuel, comparé à ces mêmes conditions pour une présentation uniquement visuelle ou uniquement sonore, confirme la présence d'un effet de redondance de cible facilitateur. Ces résultats devraient encourager les concepteurs d'interfaces à présenter le contenu audio en plus du rendu visuel durant la présentation de documents multimédia simultanés.

#### 6.7.3.3 Limites du flou audiovisuel comme trait caractéristique bimodal

Nous n'avons pas eu connaissance d'autre étude s'intéressant à l'effet de la combinaison redondante de traits caractéristiques issus de deux modalités sensorielles. Notre étude serait donc unique sur cette problématique. Les résultats montrent que le trait caractéristique de netteté (obtenu en rendant flous les distracteurs) appliqué à la modalité visuelle seule conduit à des performances similaires à celles obtenues lorsqu'on applique un trait caractéristique à la fois en audio et en visuel. Il n'y a donc pas de gain supplémentaire dû à l'effet de redondance de traits caractéristiques comme on avait pu le constater dans une combinaison de deux traits caractéristiques visuels [Krummenacher *et al.* 2001].

Au contraire, appliquer le trait caractéristique uniquement à la modalité auditive, en laissant une présentation visuelle neutre (comme dans la combinaison $\mathbf{AV^{neut}}$), réduit considérablement l'effet d'attraction observé dans la condition audio seule et dans la condition bimodale avec le trait appliqué aux deux modalités. La contribution du flou audio comme indice de guidage est donc plus limitée en présence d'une présentation visuelle efficace. Cette observation est en accord avec la théorie d'*efficacité inverse* [Stein et Meredith 1993] selon laquelle l'intégration multimodale, et donc l'apport de l'audio, est plus limitée lorsque l'information initiale est suffisante.

Ces résultats ont un impact direct pour les concepteurs d'interfaces de recherche de documents multimédia basées sur une présentation simultanée des documents. En effet, puisqu'ajouter un rendu audio est bénéfique tandis qu'une distorsion auditive ne l'est pas si une technique à base de distorsion visuelle est déjà utilisée, nous recommandons aux concepteurs d'IHM d'associer une composante auditive à leurs outils de présentation sans toutefois chercher à guider la recherche auditive. Cela représente ainsi une économie de temps en amont puisque les concepteurs n'ont plus à concevoir l'effet de distorsion sonore le plus adapté, et une économie de temps de calcul lors du traitement temps réel des signaux audio.

#### 6.7.3.4 Conception d'une lentille avec profondeur de champ sémantique audiovisuelle

Bien que dans notre étude nous n'ayons cherché à augmenter la saillance que d'un seul objet (net) à la fois, on peut imaginer une généralisation dans laquelle plusieurs documents sont sélectionnés et laissés nets parmi l'ensemble des documents disponibles de la collection. Ainsi une technique focus+contexte pourrait être une lentille de « défloutage » par laquelle les documents à l'intérieure seraient nets tandis que tous ceux à l'extérieur seraient flous. La position de la lentille serait contrôlable par le curseur de la souris de sorte que l'utilisateur puisse choisir sur quels documents se concentrer. La sélection de documents nets serait alors plus saillante et donc plus facilement identifiable que le reste flou de la collection, facilitant ainsi l'attention sélective de l'utilisateur. On peut alors parler de ***profondeur de champ sémantique audiovisuelle*** en référence à la technique focus+contexte de Kosara [Kosara

2001]. Cette technique audiovisuelle pourrait alors prendre en compte le contenu audio des documents multimédia tout en permettant à l'utilisateur de balayer du regard plusieurs de ces documents. Il reste évident que cette technique est plus profitable pour des collections de documents où le contenu sonore est important pour la compréhension du document. Si l'on reprend nos remarques précédentes concernant l'intérêt de l'audio mais la faiblesse du flou audio par rapport au flou visuel, une recommandation pour les concepteurs de telles lentilles de « défloutage » serait de n'appliquer le flou qu'à la composante visuelle des documents mais de laisser le contenu audio intact, sous réserve que le nombre de sources sonores audibles soit limité, par exemple en rendant les sources à l'extérieur de la lentille silencieuses.

## 6.8 Extension de l'étude à des vidéos : une expérience en cours

Les travaux que nous avons menés ici afin de définir de nouveaux paramètres pour augmenter la saillance d'une vidéo présentée parmi plusieurs autres vidéos ne constituent qu'une première étape dans une recherche sur le long terme. D'autres expérimentations sont nécessaires avant de pouvoir intégrer ces paramètres de façon efficace dans des interfaces d'exploration de collections multimédia.

En particulier, nous devons vérifier que le paramètre de flou audiovisuel que nous venons de proposer est généralisable à des vidéos. Nos résultats ont montré qu'un effet de flou audiovisuel appliqué à des distracteurs permet d'attirer l'attention sur une cible nette lorsque tous les stimuli sont des stimuli audiovisuels, constitués d'une image fixe et d'un mot prononcé. Notre hypothèse est que cet effet d'attraction s'observe également sur des stimuli audiovisuels plus complexes et plus réalistes, comme des vidéos constituées de séquences d'images (au contraire d'images fixes) et de phrases plus longues, et donc que cette attraction peut servir à la conception d'une stratégie de présentation dans une interface bimodale pour la navigation dans un ensemble de vidéos. Afin de valider cette hypothèse, nous avons cherché à mettre en place une évaluation perceptive reprenant un protocole expérimental similaire à celui employé pour l'expérience 3 mais en utilisant cette fois des vidéos. La mise en place d'une telle expérience nécessite deux choses : d'une part un corpus de vidéos et, d'autre part, un programme d'évaluation permettant de lire de façon synchronisée le contenu audio et le contenu visuel de 6 vidéos en parallèle, tout en permettant d'appliquer éventuellement sur l'une ou l'autre des modalités un effet de flou. Cette expérience est en cours et nous n'avons pu la terminer pour des questions techniques sur le développement du programme d'évaluation.

### 6.8.1 Préparation d'un corpus de vidéos

Concernant la conception d'un corpus de vidéos, les vidéos de type clip musical utilisées au chapitre 4 pourraient être réutilisées dans cette expérience. Cependant, pour comparer nos résultats avec ceux obtenus dans l'expérience 3 sur des stimuli simplifiés liés au langage (un mot écrit + le même mot prononcé), nous avons jugé préférable d'utiliser des vidéos avec de la parole non chantée. De plus, les bandes son des stimuli du chapitre 4 étaient constituées de nombreux instruments simultanés. Nous préférons, pour cette étude perceptive, utiliser des stimuli où une seule source sonore (la voix d'une personne) est présente. C'est pour ces raisons que nous avons créé notre corpus de vidéos à partir d'interviews. Nous avions à notre disposition 14

interviews d'avocats sous forme de séquences vidéos réelles enregistrées lors du forum juridique de l'université Paris 1 Panthéon-Sorbonne et disponibles sur Internet [5]. Ces vidéos ont été choisies également car elles présentent un cadre et un arrière-plan visuels très similaires (même salle et même cadrage lors de l'enregistrement), mais les personnes interviewées sont très distinctes (vêtements, visage, voix) comme on peut le voir sur la figure 6.17. Pour chacune des interviews, nous avons sélectionné plusieurs extraits assez courts, tous entre 1 et 3 s (c'est-à-dire entre 25 à 75 images par vidéos), avec une phrase ou un morceau de phrase présentant une unité syntaxique compréhensible hors de son contexte (comme « il faut avoir de bons réflexes de base » ou « ça reste un examen malgré tout »), et contenant entre 7 et 14 syllabes. Cela assure une homogénéité dans la durée des stimuli et dans la vitesse d'élocution. Finalement, avant d'utiliser ces extraits vidéos dans une expérience où plusieurs sources sont présentées concurremment, avec un effet de masquage audio, nous voulions nous assurer que les extraits sélectionnés étaient objectivement intelligibles lorsqu'écoutés individuellement, sans masquage audio et sans support visuel. Nous avons opté pour un test objectif. Nous avons d'abord annoté manuellement le texte prononcé dans chaque extrait. Ensuite, les différents extraits étaient automatiquement transcrits sous forme textuelle par un système de transcription automatique avec l'hypothèse que, si une machine est capable de reconnaître l'extrait, alors un être humain parlant la langue devrait avoir encore plus de facilité à le faire. Nous avons donc comparé les annotations manuelles et les annotations automatiques du système de reconnaissance automatique intégré dans le projet CHIL (*Computers in the Human Interaction Loop*) du LIMSI [Potamianos *et al.* 2009]. Nous avons décidé de retirer tous les extraits pour lesquels l'annotation automatique présentait une confusion (un mot au lieu d'un autre), une addition ou une omission de mots. Après ce retrait, nous avons finalement conservé 6 extraits par interview. Au final, le corpus de vidéos que nous avons constitué contient 84 (6×14) vidéos.

**Figure 6.17.** Une image extraite de chacune des 14 interviews du corpus.

### 6.8.2 Protocole expérimental

Lorsque l'implémentation logicielle sera mise en place, nous pourrons vérifier qu'un objet net parmi des objets flous attire l'attention même s'il s'agit d'un document multimédia complexe, avec des séquences d'images et de l'audio. Le protocole expérimental auquel nous avons pensé

5. Ces vidéos à but pédagogique ont été obtenues grâce à Marina Kugler de l'équipe Edition et Production Numérique du service TICe de l'Université Paris 1. Ces vidéos sont disponibles à l'adresse : http://epi.univ-paris1.fr/34496549/0/fiche___pagelibre/&RH=epi-182-itwpro-2009&RF=epi-182

reprend les grandes lignes du protocole expérimental utilisé dans l'expérience 3 précédente, notamment au niveau des variables indépendantes utilisées : modalités et conditions de congruence. Six stimuli seront donc présentés simultanément, avec un effet de flou appliqué à l'un, tous ou aucun des stimuli selon la condition de congruence. Cependant, du fait que nous utilisons des stimuli réalistes, la tâche que nous avions employée dans l'expérience précédente ("6 ou 10 ?") n'est plus valable. Nous proposons de reprendre la tâche utilisée pour l'évaluation de la lentille grossissante (chapitre 4) qui consiste à d'abord regarder entièrement une vidéo présentée individuellement (la cible) puis, après avoir pris connaissance de cette vidéo, à la retrouver parmi les 6 vidéos possibles et à la sélectionner par un clic en plaçant le curseur de la souris dessus. Comme cette tâche repose sur du pointage, il n'est pas envisageable de tester une condition Audio-seul puisqu'il faudrait soit ajouter une représentation visuelle à l'écran pour pouvoir définir les zones cliquables, soit utiliser une tâche de pointage manuelle. Or, si la spatialisation sonore facilite la ségrégation, la localisation auditive reste, elle, limitée, ce qui rend difficile le pointage manuel en audio. De même, la recherche d'un document vidéo complètement silencieux est en fait peu réaliste, surtout dans ce cas de figure où les documents sont des interviews dont l'information sémantique est essentiellement portée par la modalité auditive. Nous nous proposons donc de limiter, ce qui réduit aussi le temps d'expérience, le nombre de conditions de modalités à trois : audiovisuel avec le flou appliqué aux deux modalités **AV**, ou sur l'audio seulement $\mathbf{AV^{neut}}$ ou encore sur le visuel seulement $\mathbf{A^{neut}V}$.

De plus, compte tenu des résultats obtenus dans les expériences précédentes de ce chapitre, si les résultats sont généralisables à des vidéos d'interview alors nous pouvons émettre comme hypothèses que :

- l'utilisateur doit être attiré par la vidéo nette qu'il s'agisse de la cible ou non. Ainsi, pour les conditions où la vidéo nette est la cible (conditions **Congruentes**), les temps de réponse doivent être plus courts que pour des conditions **Neutres** où tout est net. Au contraire, pour les conditions où la vidéo nette est un distracteur mais que la cible est floue (conditions **Incongruentes**), les temps de réponse doivent être plus longs que pour la condition où tout est flou (conditions **Dégradées**) ;
- une vidéo cible sera retrouvée plus difficilement, c'est-à-dire moins correctement retrouvée et après un temps de recherche plus long, si elle est floue (conditions **Dégradée** ou **Incongruente**) que si elle est nette (conditions **Neutre** ou **Congruente**) ;
- l'effet d'attraction, mesuré par les temps de réponse, devrait être équivalent dans une condition audiovisuelle où les deux modalités sont floutées (**AV**) et dans une condition où seul le visuel est flouté ($\mathbf{A^{neut}V}$).

Il est cependant possible que cette dernière hypothèse dépende aussi de l'importance du contenu sémantique de chaque modalité. Dans les expériences précédentes, nous utilisions des stimuli image fixe+mot prononcé où les deux composantes, audio et visuelle, apportaient une information permettant de différencier la cible des distracteurs. Notre corpus de stimuli vidéos, constitué d'interviews, nous permet cependant de limiter l'apport de la modalité visuelle car les six vidéos présentées simultanément lors d'un essai peuvent :

- être extraites d'interviews différentes, auquel cas la reconnaissance visuelle de l'interviewé (reconnaissance faciale) qui est une information visuelle forte permet d'identifier la cible ;

- être extraites de la même interview, auquel cas seuls les mouvements de la tête et des mains peuvent aider visuellement à identifier la cible ce qui constituent une aide visuelle faible surtout si les vidéos sont floues. La modalité auditive est alors beaucoup plus déterminante que la modalité visuelle pour identifier la cible.

Il serait intéressant de décliner le protocole expérimental selon ces deux possibilités pour évaluer si, dans le cas où la modalité auditive est quasiment la seule à permettre l'identification, le flou conduit aussi à un effet d'attraction plus important s'il est appliqué sur la modalité visuelle plutôt que sur la modalité auditive.

### 6.8.3 Implémentation logicielle

Finalement, la dernière étape nécessaire pour mettre en place complètement cette nouvelle expérience est de développer un programme pour lire les contenus audio et visuel, synchronisés, de 6 vidéos en parallèle avec éventuellement un effet de flou. L'utilisation de séquences d'images plutôt que d'images fixes compliquent nettement l'implémentation car il faut plus de ressources, notamment en ce qui concerne la mémoire vive et le traitement graphique, pour à la fois charger l'ensemble des stimuli à présenter simultanément et appliquer l'effet de distorsion. La résolution de ce problème est d'autant plus compliquée si l'on considère que les traitements doivent être appliqués de manière interactive et en temps réel. Sans succès, nous avons tenté à deux reprises, sous *Virtual Choreographer* puis sous *OpenCV*, d'implémenter le programme dédié à l'expérience dans les temps impartis. Les difficultés d'implémentation que nous avons rencontrées soulèvent une nouvelle question : « quels outils techniques peut-on utiliser pour mettre en place une architecture logicielle qui permettent d'afficher simultanément de nombreuses vidéos et d'appliquer en plus des effets de distorsion en temps réel ? ». Cette question représente un réel challenge et nécessite un travail de développement et d'optimisation important que nous n'avons pas eu le temps de fournir ici.

## 6.9 Conclusion et perspectives de l'étude

Cette étude avait pour but de répondre au manque de stratégies de présentation sonores en proposant de nouveaux paramètres acoustiques utilisables dans les interfaces zoomables pour mettre en avant l'information pertinente. Nous nous sommes appuyée sur un paramètre visuel qui avait déjà fait ses preuves pour faciliter les tâches de recherche visuelle et nous avons procédé par analogie pour l'étendre à la modalité auditive. Nous avons alors proposé et défini un effet de flou audio pour augmenter la saillance d'une source sonore nette présentée parmi plusieurs sources sonores floues concurrentes. Nous avons également défini un effet de flou audiovisuel par combinaison des flous visuels et sonores. Une série d'expériences a ensuite confirmé que les flous audio et audiovisuels obtenus peuvent être utilisés pour attirer l'attention sur un objet net parmi des distracteurs flous.

Cette étude comparait également chacune des deux tâches de recherche unimodale à une tâche de recherche bimodale en proposant différentes combinaisons audiovisuelles. Une présentation bimodale améliore significativement les performances en termes de temps de réponse. Cela confirme l'intérêt de l'audio pour des interfaces destinées au multimédia, intérêt que nous avions déjà observé dans les études présentées précédemment notamment sur les

méthodes Pan&Zoom et lentille Fisheye audiovisuelles (chapitre 4).

En revanche, la contribution du flou audio comme indice de guidage s'est révélée limitée en présence d'une présentation visuelle cohérente. Ces résultats montrent encore une fois qu'il n'est pas nécessaire d'appliquer le même traitement de distorsion aux deux modalités. Toutefois, alors que dans le cas de l'étude sur les méthodes Pan&Zoom et lentille Fisheye, nous avions obtenu qu'il était préférable de laisser le rendu visuel sans distorsion particulière et, au contraire, d'appliquer une forte distorsion au rendu audio, nous avons trouvé ici que la déformation du rendu audio n'est pas utile. Ces résultats laissent supposer que l'optimisation de la combinaison audiovisuelle et des méthodes de distorsion proposées dans chaque modalité, dépend des paramètres audio et visuels utilisés pour la distorsion. Les deux études sont cependant cohérentes dans le sens où elles ont toutes les deux montré que, si la stratégie de présentation visuelle est déjà efficace, l'ajout de la modalité auditive a un effet moindre. Dans l'étude sur la lentille et le Pan&Zoom, nous avions conclu que, pour augmenter le gain, il fallait augmenter la distorsion. Ici il s'avère que la distorsion auditive n'améliore pas les performances dans le cas d'une présentation audiovisuelle. Une étude plus poussée devrait être menée pour conclure sur cette question.

Au-delà de l'utilisation première que l'on cherche à en faire pour de l'exploration de collections multimédia, le floutage de distracteurs en présence d'une cible laissée nette peut être mis à profit dans d'autres types d'application où il est important d'attirer l'attention de l'utilisateur sur un élément d'information particulier. On peut alors penser à la publicité. Cette idée fait d'ailleurs l'objet de nouveaux projets de recherche sur l'utilisation de sons spécifiques pour guider l'attention visuelle sur un objet particulier dans un rayon de supermarché [Knöferle et Spence 2012].

Les résultats obtenus ont aussi montré que l'application d'un effet de flou sur des sources sonores concurrentes de moindre intérêt augmente l'intelligibilité de la source sonore nette. Cette mise en relief automatique de certains objets permet la création d'interfaces sonores basées sur la présentation de plusieurs sons simultanés, soit destinées à l'exploration de collections sonores ou multimédia, soit destinées à du travail collaboratif, comme la vidéoconférence, ou à des applications militaires dans lesquelles plusieurs locuteurs peuvent s'exprimer en même temps mais où seul le discours de l'un d'entre eux est pertinent à un instant donné.

Dans l'idée de développer par la suite des interfaces destinées à la recherche de vidéos, nous devons cependant vérifier que les résultats, obtenus dans cette expérience avec des stimuli sonores courts et des images fixes, sont généralisables à d'autres stimuli. Une expérimentation est en cours afin de vérifier que l'effet d'attraction involontaire observé sur des images fixes est conservée lorsque l'on utilise des vidéos. Elle reprend un protocole expérimental s'appuyant sur les variables de l'expérience 3 de cette étude, avec différentes combinaisons de modalité et les quatre conditions de congruence, mais utilise des stimuli vidéos, constitués de séquence d'images (au contraire d'images fixes) et de sons plus longs associés.

Chapitre 7

# Conclusion générale

*Ce qui barre la route fait faire du chemin.*
Jean de La Bruyère

**Sommaire**



Nous avons présenté, tout au long de ce manuscrit, plusieurs travaux menés pour fournir à l'utilisateur des stratégies de présentation audio et audiovisuelles dédiées à l'exploration de grandes collections multimédia.

Alors que les études antérieures sur les interfaces d'accès aux collections multimédia se consacraient au développement de stratégies exclusivement visuelles, ou, dans de rares cas, exclusivement auditives, nous nous sommes proposée de développer des outils qui allient les deux modalités. Cela permet de profiter notamment de la modalité auditive dans le cas où, pour certains documents multimédia, la composante sonore est aussi, voire plus, importante que la composante visuelle. De plus, les stratégies de présentation décrites dans ce manuscrit sont basées sur une distorsion de paramètres perceptifs qui permettent de mettre en valeur certains documents présentés parmi plusieurs, ou de hiérarchiser les différents documents par ordre d'importance. L'étude de ces paramètres nous a permis de concevoir les stratégies de présentation en fonction des capacités perceptives et attentionnelles des utilisateurs. Les différentes contributions apportées par cette thèse pour répondre à notre problématique sont récapitulées dans la section 7.1.

Les propositions de stratégies audio ou audiovisuelles proposées ont été évaluées soit par des études perceptives, soit par des études d'utilisabilité. Ces études ont confirmé l'intérêt d'une présentation auditive pour améliorer le ressenti utilisateur, mais ont révélé, en termes de performances et de temps d'exécution, des limites qui demandent des réglages supplémentaires pour tirer un profit maximal de l'ajout d'audio. Nos réflexions concernant ces réglages, ainsi que d'autres pistes de recherche pour améliorer les interfaces dédiées à l'exploration de grandes collections de documents multimédia, sont présentées dans la section 7.2.

## 7.1 Contributions de la thèse

Les stratégies de présentation auxquelles nous nous sommes intéressée dans ce manuscrit reposent sur une présentation simultanée des documents et jouent sur la notion de niveau de

détail et de saillance perceptive pour mettre en avant l'information pertinente. Une information importante est plus détaillée, et elle est présentée de sorte à être perçue comme plus saillante, car la notion de saillance fait référence à la capacité d'un objet à attirer l'attention. Ces stratégies s'appuient donc sur un ensemble de paramètres perceptifs qui permettent d'augmenter la saillance et le niveau de détail subjectif. L'attention de l'utilisateur est alors dirigée vers l'information mise en avant ce qui facilite la recherche d'un document parmi plusieurs. L'analyse de l'état de l'art a montré que, sur le plan visuel, les paramètres de taille, de couleur et de netteté sont largement exploités en IHM. Dans les interfaces sonores, ce sont principalement les paramètres de volume et de position sonore qui sont exploités.

Le premier objectif de la thèse était de proposer des stratégies de présentation audiovisuelles. Pour cela, nous avons, dans un premier temps, repris des paramètres audio et visuels déjà exploités, à savoir la taille visuelle et le volume sonore, et nous avons combiné des stratégies de présentation unimodales utilisant ces paramètres, pour en faire des stratégies multimodales. Nous avons défini, au chapitre 4, un modèle pour combiner ces stratégies unimodales. À partir de ce modèle, nous avons implémenté deux stratégies multimodales que nous avons testées sur une tâche de recherche de vidéos où la vidéo cible était présentée simultanément à 99 autres vidéos. La première méthode, dite Pan&Zoom, combinait un rendu visuel homogène où seuls quelques objets étaient affichés, tous avec la même taille, et un rendu audio avec une faible distorsion de volume, de sorte que les sons attachés aux vidéos affichées aient aussi un volume homogène. En revanche, la seconde méthode, dite Fisheye Lens, combinait un rendu visuel avec une distorsion sur la taille des objets et un rendu audio avec une distorsion prononcée sur le volume sonore.

Dans un second temps, nous nous sommes proposée de définir et d'étudier d'autres paramètres audio et visuels pour augmenter la saillance d'un objet audiovisuel. Ainsi, au chapitre 6, nous avons défini un paramètre de flou audio obtenu par analogie algorithmique avec le flou visuel par un filtrage passe-bas. Combinant les flous de chaque modalité, nous avons aussi défini un flou audiovisuel. Une série d'expérimentations a été menée pour vérifier qu'une cible nette attirait l'attention parmi des distracteurs flous, quelle que soit la modalité employée pour la présentation des stimuli. Les deux premières expériences ont ainsi permis de comparer différents niveaux de flou visuel puis différents niveaux de flou audio. Une phase de calibration multimodale a été employée pour sélectionner, dans chaque modalité, un niveau de flou qui conduisait à des performances du même ordre de grandeur. Une dernière expérience a alors permis d'évaluer l'apport de chaque modalité dans la présentation multimodale et dans le flou audiovisuel. Pour cette série d'expériences, nous avons proposé un protocole expérimental original basé sur une comparaison des performances obtenues pour différentes conditions de congruence selon que le paramètre augmentant la saillance (netteté) est appliqué à la cible ou à un des distracteurs présents. Pour stimuli, nous avons employé des paires audiovisuelles correspondantes de mots écrits et parlés. Les résultats obtenus ont confirmé que les paramètres de flou audio et audiovisuel permettent d'augmenter la saillance d'une source sonore ou audiovisuelle nette, alors mise en avant, et ainsi d'attirer de façon involontaire, donc sans effort, l'utilisateur vers celle-ci. Ces paramètres peuvent donc être utilisés pour étendre, à l'audio et à l'audiovisuel, la technique de profondeur de champ sémantique qui consiste à guider l'attention de l'utilisateur vers les documents nets.

Les études de ces deux chapitres ont confirmé qu'il est possible de définir des stratégies de présentation audiovisuelles facilitant l'accès à la fois au contenu sonore et au contenu visuel de plusieurs documents simultanés.

Le second objectif de la thèse était de vérifier l'apport d'une présentation audiovisuelle par rapport à une présentation uniquement visuelle. De nombreux travaux antérieurs mentionnaient l'avantage de la multimodalité. Dans l'étude sur les sons d'environnements, décrite au chapitre 5, les participants devaient indiquer la catégorie « *vivante* » ou « *non vivante* » de sons masqués par du bruit en présence d'images congruentes ou non. Cette étude a confirmé l'avantage de la multimodalité dans le cas où une des modalités, ici la modalité auditive, est dégradée. L'autre modalité permet alors de compenser, au moins partiellement, les effets négatifs de la dégradation.

Les autres expérimentations que nous avons menées au cours de cette thèse ont également confirmé les avantages de la multimodalité sur les stratégies de présentation que nous avons proposées pour la présentation de plusieurs documents audiovisuels simultanés. Ainsi dans l'étude sur le flou au chapitre 6, on observe un gain pour une présentation bimodale, à la fois sur l'identification des objets (mots écrits+parlés) et sur les temps de réponse. De plus, lors de l'étude d'utilisabilité des méthodes Fisheye et Pan&Zoom pour la recherche d'un clip vidéo parmi 100, présentée au chapitre 4, les participants ont apprécié la possibilité de se référer à des informations complémentaires apportées par l'audio. Ces résultats liés au ressenti utilisateur sont encourageants et invitent à poursuivre les recherches pour le développement de stratégies de présentation multimodales dédiées à la recherche de vidéos comme nous le présentons ci-dessous en sections 6.8 et 7.2. Les performances musurées dans cette étude sur les vidéos musicales n'ont cependant pas montré d'amélioration de la présentation audiovisuelle par rapport à une présentation uniquement visuelle, sans toutefois en montrer d'effet négatif. Le fait qu'il n'y ait pas de gain en termes de temps de réponse indique aussi des limites de la multimodalité et la nécessité d'approfondir les recherches pour optimiser le rendu multimodal et tirer un profit maximal de chacune des modalités.

Concernant l'optimisation de la combinaison audiovisuelle, une des questions que nous nous posions concernait la nécessité d'employer une distorsion dans les deux modalités lorsque les stratégies audio et visuelle employées initialement sont des méthodes à bases de distorsion.

À cette question, deux études différentes, présentées dans les chapitres 4 et 6, révèlent qu'il n'est pas nécessaire d'utiliser une distorsion dans chacune modalité. Dans la première étude, qui examine des stratégies audiovisuelles telles que la lentille grossissante et la méthode Pan& Zoom sur des vidéos, les résultats objectifs et subjectifs indiquent une nécessité de distordre fortement le rendu audio pour limiter le nombre de sources sonores jouées simultanément, mais de laisser homogène, c'est-à-dire sans distorsion, le rendu visuel. Dans la seconde étude, qui examine le flou audiovisuel comme paramètre facilitateur de recherche, les résultats indiquent que si la multimodalité est bénéfique, une combinaison multimodale où le visuel et l'audio sont floutés est équivalente à une combinaison multimodale où seul le visuel est flouté. Dans cette étude, la distorsion visuelle est nécessaire pour obtenir l'effet d'attraction souhaité mais pas la distorsion sonore.

Ces deux études laissent entendre qu'il suffit d'appliquer la distorsion sur une seule des modalités. Cependant nous proposons une autre interprétation qui tient compte de la prédominance

visuelle, que nous avons notée à travers l'état de l'art et qui est aussi relevée par les résultats de nos expérimentations. Les résultats ont montré que les utilisateurs apprécient de disposer des deux modalités dans une interfaces, que pour des documents audiovisuels complexes le contenu audio peut contenir des informations sémantiquement discriminatoires pour la recherche d'un document spécifique, mais que des réglages spécifiques sur l'audio sont nécessaires pour tirer profit de cette modalité sans gêne. Au final, cela suggère que pour concevoir des interfaces d'exploration de collection multimédia, il est préférable de proposer une présentation multimodale, avec quelques recommandations supplémentaires pour les rendus audio et visuel. Pour être plus précise, le rendu audio doit limiter le nombre de sources sonores entendues simultanément. La nécessité de distordre ou non la modalité auditive semble liée à ce dernier point et à l'efficacité de la modalité visuelle. Si la stratégie visuelle employée pour augmenter la saillance ou le niveau de détail est suffisamment efficace, il ne semble pas nécessaire de distordre la modalité auditive.

Des expériences supplémentaires sont nécessaires pour valider l'une ou l'autre des interprétations, en comparant par exemple des combinaisons audiovisuelles on les deux modalités sont floutées mais le niveau de flou audio varie. Cela permettrait également une meilleure optimisation de la combinaison audiovisuelle.

## 7.2 Perspectives de recherche

Nos travaux constituent les premières recherches sur la problématique du développement de stratégies de présentation audiovisuelles pour l'exploration de collection multimédia. Bien que nous ayons répondu à plusieurs questions scientifiques liées à cette problématique, de nombreuses pistes de recherche restent encore ouvertes.

Nos travaux nous ont permis de découvrir qu'il n'est pas nécessaire, lors d'une combinaison multimodale, de combiner les effets de distorsion de chaque modalité. Cette découverte soulève différentes questions. Par exemple, d'après la conclusion du chapitre 4, nous devons vérifier qu'il est effectivement préférable de laisser un rendu visuel homogène et un rendu audio très distordu si l'on présente de très nombreuses vidéos avec pour méthode un jeu sur la taille et le volume sonore. Cette possibilité d'associer une stratégie de présentation différente pour chaque modalité laisse également penser qu'il est possible, voire préférable, d'utiliser des stratégies s'appuyant sur des paramètres tout à fait différents dans chaque modalité. On peut alors imaginer un système où le niveau de détail du visuel est rendu par la taille des objets, mais sans distorsion (méthode dite Pan&Zoom), tandis que pour l'audio le rendu est basé sur du flou audio (distorsion fréquentielle). Sur ce point, il est d'ailleurs important de comprendre que nous n'avons utilisé que deux paramètres pour la conception de stratégies de présentation sonores ou audiovisuelles, la taille / volume et la netteté. Plusieurs autres paramètres visuels sont utilisés en IHM, et l'on peut imaginer que d'autres paramètres sonores pourraient être utilisées dans des interfaces sonores basées sur une présentation simultanée des informations. Par exemple, nous avons déjà suggéré au chapitre 6 que la réverbération pourrait servir de trait caractéristique auditif sous forme de flou cinétique audio, comme le filtrage fréquentiel sous forme de flou statique audio. Au final, cela laisse entendre que le nombre de combinaisons de stratégies audio et visuelles envisageables est très large.

La création d'interfaces d'exploration de collection multimédia nécessite aussi d'élargir nos recherches à d'autres domaines d'études. En effet, nous avons limité nos travaux à l'étude de l'interaction en sortie (communication de l'ordinateur vers l'utilisateur). L'interaction en entrée, c'est-à-dire la façon dont l'utilisateur peut commander l'interface, naviguer ou envoyer des requêtes à l'ordinateur, n'a pas du tout été abordée. D'ailleurs, la comparaison entre plusieurs stratégies de présentation au chapitre 4 a pu être biaisée par l'utilisation d'un type de contrôle qui correspondait mieux à l'une des deux techniques. Il est évident que, pour la création d'une interface d'exploration de bases de données, l'interaction en entrée est extrêmement importante. De nombreuses études se penchent d'ailleurs sur la question, par exemple [Gutwin 2002; Serrano 2010], que ce soit au niveau des dispositifs d'entrée (clavier, souris, souris 3D, gants, captation visuelle, surface tactile, etc.) ou des méthodes de déplacement et d'interaction (saisie textuelle, clic, double-clic, glisser-déposer, difficultés de pointage, interaction bi-manuelle, communication naturelle). À chaque stratégie de présentation correspond une méthode de contrôle optimale. Il est important d'étudier conjointement le contrôle et la présentation (qui sert ici de retour ou *feedback*). Cela est d'autant plus important si l'on pense à l'exploration de collection de documents pour un travail collaboratif où plusieurs personnes seraient amenées à explorer la même collection, mais pourquoi pas à voir et à entendre des documents différents, comme, par exemple, sur un équipement de type table interactive (*tabletop*).

Enfin, un autre domaine de recherche pourrait être pris en compte pour optimiser les techniques de présentation de données, à savoir le domaine de la recherche d'information. Des travaux antérieurs avaient été menés pour associer des stratégies de présentation audio (***SonicBrowser***) à un système d'analyse automatique et d'indexation (***MARSYAS***) dans l'étude [Brazil *et al.* 2002]. De la même manière, en tenant compte des critères sur lesquels se basent les analyses de vidéos [Huurnink *et al.* 2012], et en considérant que les documents de la collection multimédia à traiter sont organisés, nous pourrions proposer une disposition des documents audiovisuels, et des outils à base de distorsion, plus adaptés. Inclure la possibilité de faire une recherche par requête (mots-clés ou similarité) permettrait de fournir et d'étudier un système complet où les différentes composantes de l'interface, de l'utilisateur au rendu en passant par le contrôle et la collection elle-même, sont prises en compte conjointement pour que chacune amplifie les bénéfices des autres composantes.

# Références bibliographiques


Apperley, M. et Spence, R. (1982). A bifocal display technique for data presentation. *In Eurographics 1982*, pages 27–43. (Cité page 34)

Arnold, P. et Hill, F. (2001). Easy to hear but hard to understand : a lipreading advantage with intact auditory stimuli. *British Journal of Psychology*, 92:339–355. (Cité page 50)

Arons, B. (1992). A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 12:35–50. (Cité page 37)

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology : Human Perception and Performance*, 19:250–267. (Cité pages 93 et 94)

Ballas, J. A. et Mullins, R. T. (1991). Effects of context on the identification of everyday sounds. *Human Performance*, 4(3):199–219. (Cité pages 97, 98, et 118)

Begault, D. R. (1994). *3D Sound for Virtual Reality and Multimedia.* MA : Academic Press Professional. (Cité pages 21, 42, et 46)

Berkhout, A. J., de Vries, D. et Vogel, P. (1993). Acoustic control by wave field synthesis . *Acoustical Society of America Journal*, 93:2764–2778. (Cité page 45)

Bertelson, P. et Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics*, 29(6):578–584. (Cité pages 48 et 49)

Best, V., Ozmeral, E. J. et Shinn-Cunningham, B. G. (2007). Visually-guided attention enhances target identification in a complex auditory scene. *Journal of the Association for Resarch in Otolaryngology*, 8:294–304. (Cité page 52)

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M. et Rzeczkowski, C. (0001). Standardization of a test of speech perception in noise. *Journal of speech and hearing research*, 27(1):32 – 48. (Cité page 96)

Blauert, J. (1997). *Spatial Hearing. The Psychophysics of Human Sound Localization.* MIT Press. (Cité page 124)

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345. (Cité page 112)

Bolia, R. S., D'Angelo, W. R. et McKinley, R. L. (1999). Aurally aided visual search in three-dimensional space. *Human Factors*, 41:664–669. (Cité page 52)

Bolognini, N., Frassinetti, F., Serino, A. et Làdavas, E. (2005). "acoustical vision" of below threshold stimuli : interaction among spatially converging audiovisual inputs. *Experiemental Brain Research*, 160:273–282. (Cité page 50)

Bouchara, T. (2008). Le “scenemodeler” : des outils pour la modélisation de contenus multimédias interactifs spatialisés. *In* GMEA-AFIM, éditeur : *13ème Journées d'Informatique Musicale (JIM'08)*, Albi, France. (Cité page 72)

Bouchara, T., Giordano, B. L., Frissen, I., Katz, F. et C., G. (2010a). Effect of signal-to-noise ratio and visual context on environmental sound identification. *In 128th convention of the Audio Engineering Society (AES128th)*, London, UK. (Cité page 6)

Bouchara, T., Guastavino, C., Katz, B. et Jacquemin, C. (2010b). Audio-visual renderings for multimedia navigation. *In 16th International Conference on Auditory Display (ICAD-2010)*, pages 245–252, Washington, DC. (Cité page 5)

Bouchara, T. et Katz, B. F. (2012). Redundancy gains in audio-visual search. *Seeing and Perceiving*, 25:181–181(0). (Cité page 6)

Bouchara, T., Katz, B. F. et Jacquemin, C. (2012). Guidage attentionnel à base de flou audiovisuel pour la conception d'interfaces multimodales. *In Conférence francophone sur l'interaction homme machine et l'ergonomie (Ergo'IHM2012)*, Biarritz, France. ACM. (Cité page 6)

Boyne, S.and Pavlovic, N., Kilgore, R. et Chignell, M. (2005). Auditory and visual facilitation : Cross-modal fusion of information in multi-modal displays. *In Visualisation and the Common Operational Picture. Meeting Proceedings RTO-MP-IST-, 043. paper 19.*, pages 19–1 19–4, Neuilly-sur-Seine, France : RTO. (Cité page 26)

Bradley, J., Reich, R. et Norcross, S. (1999). On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility. *Journal of the Acoustical Society of America.*, 106:1820–1828. (Cité page 124)

Brazil, E. (2003). Investigation of multiple visualisation techniques and dynamic queries in conjunction with direct sonification to support the browsing of audio resources. Mémoire de D.E.A., University of Limerick. (Cité page 22)

Brazil, E., Fernstroem, M., Tzanetakis, G. et Cook, P. (2002). Enhancing sonic browsing using audio information retrieval. *In* Nakatsu, R. et Kawahara, H., éditeurs : *ICAD International Conferences of Auditory Display*, Kyoto, Japan. Advanced Telecommunications Research Institute (ATR), Kyoto, Japan, Advanced Telecommunications Research Institute (ATR), Kyoto, Japan. (Cité page 165)

Brazil, E. et Fernström, M. (2011a). *The Sonification Handbook*. Chapitre 13. Auditory Icons, page 325–338. Logos Verlag. (Cité pages 89 et 118)

Brazil, E. et Fernström, M. (2011b). *The Sonification Handbook*. Chapitre 14. Earcons, page 339–361. Logos Verlag. (Cité page 89)

Bregman, A. S. (1990). *Auditory scene analysis.The perceptual organisation of sound*. Cambridge, Mass. : Bradford Books. (Cité pages 21, 37, 53, 64, 129, et 146)

Brewster, S. A. (1994). *Providing a Structured Method for Integrating Non-Speech Audio into Human-Computer Interfaces*. Thèse de doctorat, University of York. (Cité page 19)

Bronkhorst, A. W. (2000). The cocktail party phenomenon : A review on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86:117–128. (Cité pages 37, 95, 124, et 129)

Brungart, D. et Simpson, B. (2005). Improving multitalker speech communication with advanced audio displays. *In New Directions for Improving Audio Effectiveness. Meeting RTO-MP-HFM-123.*, numéro paper 30, pages 30.1 – 30.18, Neuilly-sur-Seine, France. (Cité pages 40, 129, 130, et 145)

Burr, D. et Alais, D. (2006). Combining visual and auditory information. *Visual Perception, Pt 2 : Fundamentals Of Awareness : Multi-Sensory Integration And High-Order Perception*, 155:243–258. (Cité page 51)

Calvert, G., Spence, C. et Stein, B. (2004). *The Handbook of Multisensory Processes*. MIT Press. (Cité page 47)

Caramazza, A. et Shelton, J. R. (1998). Domain-specific knowledge systems in the brain : The animate-inanimate distinction. *J. Cognitive Neuroscience*, 10(1):1–34. (Cité page 117)

Carlile, S. et Schonstein, D. (2006). Frequency bandwidth and multi-talker environments. *In 120th Convention of the Audio Engineering Society*. (Cité page 146)

Carlyon, R., Cusack, R., Foxton, J. et Robertson, I. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology : Human Perception and Performance*, 27:115–127. (Cité page 37)

Carpendale, M. et Montagnese, C. (2001). A framework for unifying presentation space. *In the 14th Annual ACM Symposium on User Interface Software and Technology*, pages 61–70. (Cité page 14)

Carpendale, S. (2008). A theory of elastic presentation space. Cours accessible à l'adresse innovis.cpsc.ucalgary.ca/innovis/uploads/Courses/InformationVisualizationDetails/05elastic-presentation-part2.pdf. (Cité page 15)

Chareyron, G. (2005). *Tatouage d'images : une approche couleur*. Thèse de doctorat, Université Jean-Monnet Saint-Etienne. (Cité page 133)

Chen, Y.-C. et Spence, C. (2009). Crossmodal facilitation of visual target identification at the level of object representation by the presentation of a concomitant sound. *In European Conference on Visual Processing*. (Cité page 50)

Chen, Y.-C. et Spence, C. (2010). When hearing the bark helps to identify the dog : Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, 114(3):389 – 404. (Cité page 50)

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25(5):975–979. (Cité pages 21, 37, et 38)

Chion, M. (1990). *L'Audio-Vision*. Nathan. (Cité page 89)

Christmann, O. (2008). *Navigation dans de grands ensembles non structurés de documents visuels*. Thèse de doctorat, Ecole Doctorale IAEM Lorraine. (Cité page 12)

Cockburn, A., Karlson, A. et Bederson, B. B. (2008). A review of overview+detail, zooming, and focus+context interfaces. *ACM Computing Surveys*, 41(1). (Cité pages 12 et 13)

Couvreur, L., Bettens, F., Drugman, T., Frisson, C., Jottrand, M., Mancas, M. et Moinet, A. (2008). Audio skimming. *QPSR of the numediart research program*, 1(1). (Cité page 24)

Culling, J. F., Hodder, K. I. et Toh, C. Y. (2003). Effects of reverberation on perceptual segregation of competing voices. *Journal of the Acoustical Society of America.*, 114:2871–2876. (Cité page 124)

Cusack, R. et Carlyon, R. P. (2003). Perceptual asymmetries in audition. *Journal of Experimental Psychology : Human Perception and Performances.*, 29(3):713–725. (Cité page 40)

Cusack, R., Deeks, J., Aikman, G. et Carlyon, R. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J Exp Psychol Hum Percept Perform*, 30:643–656. (Cité page 38)

Daniel, J. (2000). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. Thèse de doctorat, Université de Paris 06. (Cité page 45)

de Bruijn, O. et Spence, R. (2000). Rapid serial visual presentation : a space-time trade-off in information presentation. *In working conference on Advanced Visual Interfaces*. (Cité page 17)

Doyle, M. C. et Snowden, R. J. (2001). Identification of visual stimuli is improved by accompanying auditory stimuli : The role of eye movements and sound location. *Perception*, 30:795–810. (Cité page 50)

Duncan, J. et Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*, 96:433–458. (Cité page 33)

Eramudugolla, R., McAnally, K. I., Martin, R. L. et andJason B. Mattingleya, D. R. I. (2008). The role of spatial location in auditory search. *Hearing Research*, 238:139–146. (Cité pages 40 et 129)

Ernst, M. O. et Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4):162 – 169. (Cité pages 47, 54, et 106)

Fernström, M. et Brazil, E. (2001). Sonic browsing : an auditory tool for multimedia asset management. *In International Conference on Auditory Display*, pages 132–135, Espoo, Finland. (Cité pages v, 22, 23, et 82)

Frederiksen, J. R. (1967). Cognitive factors in the recognition of ambiguous auditory and visual stimuli. *Journal of Personality and Social Psychology*, 7(1, Part 2):1 – 17. (Cité pages 122 et 124)

French, R. M. et Mareschal, D. (1998). Could category-specific semantic deficits reflect differences in the distributions of features within a unified semantic memory ? *In 20th Annual Conference of the Cognitive Science Society*. (Cité page 117)

Furnas, G. W. (1981). The fisheye view : a new look at structured file. Rapport technique, Bell Laboratories. LE article à citer !!! (Cité page 15)

Furnas, G. W. (1986). Generalized fisheye views. *In Conference on Human Factors in Computing Systems CHI'86*, pages 18–23. (Cité page 15)

Furnas, G. W. et Bederson, B. B. (1995). Space-scale diagrams : Understanding multiscale interfaces. *In* ACM, éditeur : *Human Factors in Computing Systems CHI '95*, pages 234–241. (Cité pages v, 13, 14, et 65)

Gaver, W. W. (1989). The sonicfinder, an interface that uses auditory icons. *Human Computer Interaction*, 4:67–94. (Cité page 89)

Gaver, W. W. (1993). What in the world do we hear ? an ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29. (Cité pages vi, 90, 94, et 100)

Gérard, Y. (2004). *Mémoire sémantique et sons de l'environnement*. Thèse de doctorat, Université de Bourgogne, Dijon, France. (Cité pages 97, 98, 100, 107, 108, 116, 117, et 118)

Gerzon, M. A. (1985). Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871. (Cité page 44)

Ghirardelli, T. G. et Scharine, A. A. (2009). *Helmet Mounted Displays- Sensation, Perception and Cognitive Issues*. Chapitre Chapter 14. Auditory-Visual Interactions, pages 599–618. U.S. Army Aeromedical Research Laboratory. (Cité page 47)

Giard, M. et Peronnet, F. (1999). Audio-visual integration during multimodal object recognition in humans : A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, 11:473–490. (Cité page 50)

Giordano, B., McDonnell, J. et McAdams, S. (2010). Hearing living symbols and non living icons : category specificities in the cognitive processing of environmental sounds. *Brain&Cognition*. (Cité pages ix, 91, 99, 100, 101, 104, 106, et 117)

Gooch, B., Sloan, P.-P. J., Gooch, A., Shirley, P. et Riesenfeld, R. (1999). Interactive technical illustration. *In I3D '99 : Proceedings of the 1999 symposium on Interactive 3D graphics*, pages 31–38, New York, NY, USA. ACM. (Cité pages v et 18)

Grabowecky, M., Sherman, A. et Susuki, S. (2012). Natural scenes have matched amplitude-modulated sounds that systematically influence visual scanning. *In International Multisensory Research Forum (IMRF 2012). Seeing and Perceiving, vol.25 (S1).* (Cité page 141)

Grant, K. W. et Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Jouurnal of Acoustical Society of America*, 108(3):1197–1208. (Cité pages 49 et 51)

Grubert, A., Krummenacher, J. et Eimer, M. (2011). Redundancy gains in pop-out visual search are determined by top-down task set : Behavioral and electrophysiological evidence. *Journal of Vision*, 11(14):1–10. (Cité page 32)

Guillaume, A., Pellieux, L., Chastres, V., Blancard, C. et Drake, C. (2004). How long does it take to identify everyday sounds. *In Tenth Meeting of the International Conference on Auditory Display (ICAD'04).* (Cité page 94)

Gutwin, C. (2002). Improving focus targeting in interactive fisheye views. *In CHI '02 : Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 267–274, New York, NY, USA. ACM Press. (Cité pages 76 et 165)

Gygi, B. (2001). *Factors in the identification of environmental sounds.* Thèse de doctorat, Departement of Psychology, Indiana University. (Cité page 91)

Gygi, B., Kidd, G. R. et s. Watson, C. (2007). Similarity and categorization of environmental sounds. *Perception and Psychophysics*, 69(6):839–855. (Cité pages 90 et 93)

Gygi, B., Kidd, G. R. et Watson, C. S. (2004). Spectral- temporal factors in the identification of environmental sounds. *Journal of the Acoustical Society of America*, 115(3):1252–1265. (Cité pages 90, 91, et 93)

Gygi, B. et Shafiro, V. (2007a). Effect of audiory context on the identification of environmental sounds. *In 19th International COngress on Acoustics.* (Cité pages 96, 98, et 118)

Gygi, B. et Shafiro, V. (2007b). Environmental sound research as it stands today. *In Proceedings on Meetings on Acoustics*, volume 1, pages 1 –18. (Cité pages 89, 90, 91, et 93)

Gygi, B. et Shafiro, V. (2009). From signal to substance and back : Insights from environmental sound research to auditory display design. *In 15th International Conference on Auditory Display*, pages 1–12, Copenhagen, Denmark. (Cité pages 90 et 118)

Gygi, B. et Shafiro, V. (2011). The incongruency advantage for environmental sounds presented in natural auditory scenes. *J Exp Psychol Hum Percept Perform.*, 37(2):551–565. (Cité pages 94, 96, 98, 116, et 118)

Harrie, L., Sarjakoski, L. T. et Lehto, L. (2002). A variable-scale map for small-display cartography. *In Symposium on Geospatial Theory, Processing and Applications.* (Cité page 66)

Hawley, M., Litovsky, R. et Culling, J. (2004). The benefit of binaural hearing in a cocktail party : Effect of location and type of interferer. *Journal of the Acoustical Society ofAmerica*, 115:833–843. (Cité page 95)

Healey, C. G., Booth, K. S. et Enns, J. T. (1995). Visualizing real-time multivariate data using preattentive processing. *ACM Transactions on Modeling and Computer Simulation*, 5(3):190–221. (Cité page 34)

Healey, C. G. et Enns, J. T. (2011). Attention and visual memory in visualization and computer graphics. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*. (Cité page 30)

Heise, S., Hlatky, M. et Loviscach, J. (2008). Soundtorch : Quick browsing in large audio collections. *In 125th convention of the AES*, San Francisco, CA, USA. (Cité pages v, 22, 23, 82, 89, et 145)

Hermann, T., Hunt, A. et Neuhoff, J. G. (2011). *The sonification handbook*. Logos Publishing House. (Cité page 89)

Hollier, M. P., Rimell, A. N., Hands, D. S. et Voelcker, R. M. (1999). Multi-modal perception. *BT Technology Journal*, 17:35–46. 10.1023/A :1009666623193. (Cité pages 53 et 54)

Howard-Jones, P. et Rosen, S. (1993). The perception of speech in fluctuating noise. *Acustica*, 78:258–272. (Cité page 95)

Huurnink, B., Snoek, C. G. M., de Rijke, M. et Smeulders, A. W. N. (2012). Content-based analysis improves audiovisual archive retrieval. *IEEE Transactions on Multimedia*, 14:1166–1178. (Cité page 165)

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J. et Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics*, 72(7):1736–1741. (Cité pages 34 et 52)

Iordanescu, L., Grabowecky, M. et Suzuki, S. (2011). Object-based auditory facilitation of visual search for pictures and words with frequent and infrequent targets. *Acta Psychologica*, 137(2):252–259. (Cité pages 34, 52, et 141)

Jacquemin, C. (2004). Architecture and experiments in networked 3d audio/graphic rendering with virtual choreographer. *In Sound and Music Computing (SMC'04)*. (Cité page 65)

James, W. (1890). *The Principles of Psychology*. Chapitre 11, pages 403–404. (Cité page 30)

Katz, B., Rio, E. et Piccinali, L. (2010). LIMSI Spatialisation Engine. International Deposit Digital Number IDDN.FR.001.340014.000.S.P.2010.000.31235. (Cité page 129)

Keahey, T. A. et Robertson, E. L. (1996). Techniques for non-linear magnification transformations. *In Proceedings of the 1996 IEEE Symposium on Information Visualization (INFOVIS '96)*, INFOVIS '96, pages 38–, Washington, DC, USA. IEEE Computer Society. (Cité page 15)

Keller, P. et Stevens, C. (2004). Meaning from environmental sounds : Types of signal–referent relations and their effect on recognizing auditory icons. *Journal of Experimental Psychology : Applied.*, 10(1):3–12. (Cité page 118)

Keston (2009). John Keston's Work on Gaussian Blur for Sound Design. audiocookbook.org/sound_design/gaussian-blur/. (Cité page 122)

Kim, R., Megan A. K. Peters et Shams, L. (2011). $0 + 1 > 1$ how adding noninformative sound improves performance on a visual task. *Psychological Science*. (Cité page 52)

Kitagawa, N. et Spence, C. (2006). Audiotactile multisensory interactions in human information processing. *Japanese Psychological Research*, 48:158–173. (Cité page 47)

Knöferle, K. et Spence, C. . (2012). Product-related sounds speed visual search for products. *Seeing and Perceiving, Abstracts of the13th International Multisensory Research Forum (IMRF), University of Oxford, UK, June 19–22*, 25:193–193. (Cité page 160)

Kobayashi, M. et Schmandt, C. (1997). Dynamic soundscape : mapping time to space for audio browsing. *In CHI '97 : Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 194–201, New York, NY, USA. ACM. (Cité page 21)

Koch, I., Lawo, V., Fels, J. et Vorländer, M. (2011). Switching in the cocktail party : Exploring intentional control of auditory selective attention. *Journal of Experimental Psychology : Human Perception and Performance*, 37(4):1140–1147. (Cité page 39)

Kosara, R. (2001). *Semantic Depth of Field – Using Blur for Focus+Context Visualization*. Thèse de doctorat, Vienna University of Technology, Austria. (Cité pages v, 17, 34, 123, 126, 141, 147, et 155)

Kosara, R., Miksch, S. et Hauser, H. (2002a). Focus+context taken literally. *In GCA*. (Cité page 34)

Kosara, R., Miksch, S., Hauser, H., Schrammel, J., Giller, V. et Tscheligi, M. (2002b). Useful properties of semantic depth of field for better f+c visualization. *In Joint Eurographics – IEEE TCVG Symposium on Visualization (VisSym)*, pages 205–210. (Cité pages 34, 122, 123, 138, et 141)

Krummenacher, J., Müller, H. J. et Heller, D. (2001). Visual search for dimensionally redundant pop-out targets : Evidence for parallel-coactive processing of dimensions. *Perception & Psychophysi cs*, 63:907–917. (Cité pages 32, 130, et 155)

Kwak, H.-W., Dagenbach, D. et Egeth, H. (1991). Further evidence for a time-independent shift of the focus of attention. *Perception & Psychophysics*, 49:473–480. (Cité page 130)

Kyprianidis, J. E., Collomosse, J., Wang, T. et Isenberg, T. (2012). State of the 'ar t' : A taxonomy of ar tistic stylization techniques for images and video. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, 99. (Cité page 12)

Lamping, J., Rao, R. et Pirolli, P. (1995). A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. *In Conference on Human Factors in Computing Systems, ACM SIGCHI 95*, Denver. (Cité page 12)

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H. et Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experiemental Brain Research*, 158:405–414. (Cité page 55)

Laws, K. R. (1999). Gender affects naming latencies for living and non living things : implications for familiarity. *Cortex*, 35:729–733. (Cité page 100)

Lecolinet, E. et Pook (2002). Interfaces zoomables et âÂ« control menus âÂ» : Techniques focus+contexte pour la navigation interactive dans les bases de données. *Revue les Cahiers du numérique*, 3:191–210. (Cité page 16)

Leech, R., Gygi, B., Aydelott, J. et Dick, F. (2009). Informational factors in identifying environmental sounds in natural auditory scenes. *The Journal of the Acoustical Society of America*, 126(6):3147–3155. (Cité page 118)

Leung, Y. K. et Apperley, M. D. (1994). A review and taxonomy of distortion-oriented presentation techniques. *ACM Transaction on Computer-Human Interaction*, 1(2):126–160. (Cité pages 15 et 66)

Lewald, J. et Guski, R. (2003). C ross-modal perceptual integrationof spatiallyandtemporally disparateauditoryandvisual stimuli. *Cognitive Brain Research*, 16:468–478. (Cité pages 53 et 54)

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356–363. (Cité pages 90 et 111)

Lewis, J. W., Brefczynski, J. A., Phinney, R. E., Janik, J. J. et DeYoe, E. A. (2005). Distinct cortical pathways for processing tool versus animal sounds. *The Journal of Neuroscience*, 25(21):5148–5158. (Cité pages 90 et 117)

Li, Xand Logan, R. et Pastore, R. (1991). Perception of acoustic source characteristics : Walking sounds. *Journal of the Acoustical Society of America.*, 90(6):3036–3049. (Cité page 91)

Lieberman, H. (1997). A multi-scale, multi-layer, translucent virtual space. *In In IEEE International Conference on Information Visualization*, pages 124–131. IEEE. (Cité pages v et 16)

Lokki, T. et Gröhn, M. (2005). Navigation with auditory cues in a virtual environnement. *IEEE Multimedia*, 12(2):80–86. (Cité page 42)

Ludwig, L. F., Cohen, M. et Pincever, N. (1990). Extending the notion of a window system to audio. *Computer*, 23:66–72. (Cité page 20)

Lukas, S., Philipp, A. M. et Koch, I. (2010). Switching attention between modalities : further evidence for visual dominance. *Psychological Research*, 74:255–267. (Cité pages 57 et 150)

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J. et Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise : A bayesian explanation using high-dimensional feature space. *PLoS ONE*, 4(3):e4638. (Cité pages vi, 95, et 96)

Marcell, M. M., Borella, D., Greene, M., Kerr, E. et Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology*, 22(6):830–864. (Cité pages 89, 94, et 104)

Massaro, D. et Stork, D. (1998). Speech recognition and sensory integration. *American Scientist*, 86:236–244. (Cité page 48)

Matsumoto (2009). Akihiko Matsumoto's Work on Audio Blur. www.youtube.com/watch?v=wGttKVKUYMU. (Cité page 122)

McAdams, S. (1993). *Thinking in sound. The cognitive psychology of human audition*. Chapitre 6. Recognition of sound sources and events, pages 146–198. Oxford Science Publication. (Cité page 93)

McGookin, D. K. et Brewster, S. A. (2002). Dolphin : The design and initial evaluation of multimodal focus and context. *In International Conference on Auditory Display (ICAD'02)*, Kyoto, Japan. (Cité page 27)

McGookin, D. K. et Brewster, S. A. (2004). Understanding concurrent earcons : Applying auditory scene analysis principles to concurrent earcon recognition. *ACM Transactions on Applied Perception*, 1:120–155. (Cité page 89)

McGookin, D. K. et Brewster, S. A. (2006). Advantages and issues with concurrent audio presentation as part of an auditory display. *In 12th International Conference on Auditory Display (ICAD'06)*, London, UK. (Cité page 24)

McGurk, H. et MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748. (Cité page 47)

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, 116:374–388. (Cité page 135)

Miller, J. (1982). Divided attention : Evidence for coactivation with redundant signals. *Cognitive Psychology*, 14(2):247 – 279. (Cité page 50)

Mills, A. (1972). *Foundations of modern auditory theory*, volume 2. Chapitre Auditory Localization, pages 301–345. Academic Press. (Cité page 42)

Misdariis, N., Minard, A., Susini, P., Lemaitre, G., McAdams, S. et Parizet, E. (2010). Environmental sound perception : metadescription and modeling based on independent primary studies. *EURASIP J. Audio Speech Music Process.*, 2010:6 :1–6 :26. (Cité pages 93 et 115)

Moeck, T., Bonneel, N., Tsingos, N., Drettakis, G., Viaud-Delmon, I. et Alloza, D. (2007). Progressive perceptual audio rendering of complex scenes. *In Symposium on Interactive 3D graphics and games (I3D 2007)*. (Cité page 89)

Mynatt, E. D. (1994). Designing with auditory icons. *In 2nd International Conference on Auditory Display*. (Cité page 96)

Nicol, R. (2010). *Binaural Technology*. AES monograph. (Cité page 46)

Noisternig, M., Musil, T., Sontacchi, A. et Holdrich, R. (2003). 3d binaural sound reproduction using a virtual ambisonic approach. *In lntemafional Symposium nn Virtual Environments, Human Compuler Interfaces, and Measurement Systems*. (Cité page 46)

Occelli, V., Spence, C. et Zampini, M. (2011). Audiotactile interactions in temporal perception. *Psychonomic Bulletin & Review*, 18:429–454. 10.3758/s13423-011-0070-4. (Cité page 47)

Olivier Le Meur, T. B. (2012). Methods for comparing scanpaths and saliency maps : strengths and weaknesses. *Behavior Research Methods*. (Cité page 141)

Ozcan, E. et van Egmond, R. (2009). The effect of visual context on the identification of ambiguous environmental sounds. *Acta Psychologica*, 131(2):110 – 119. (Cité pages 50, 55, 91, 98, 116, et 118)

Parseihian, G. et Katz, B. F. G. (2012a). Morphocons : A new sonification concept based on morphological earcons. *Journal of the Audio Society of America*, 60(6):409–418. (Cité page 118)

Parseihian, G. et Katz, B. F. G. (2012b). Rapid head-related transfer function adaptation using a virtual auditory environment. *Journal of the Acoustical Society of America, Express Letters*, 131(4):10. (Cité page 46)

Pashler, H. E. (1998). *The Psychology of attention*. MIT Press. (Cité page 30)

Perrott, D., Sadralodabai, T., Saberi, K. et Strybel, T. (1991). Aurally aided visual search in the central visual field : effects of visual load and visual enhancement of the target. *Human Factors*, 33(4):389–400. (Cité page 52)

Pietriga, E. et Appert, C. (2008). Sigma lenses : Focus-context transitions combining space, time and translucenc. *In ACM CHI, Conference on Human Factors and Computing Systems*. (Cité page 15)

Pietriga, E., Appert, C. et Beaudouin-Lafon, M. (2007). Pointing and beyond : an operationalization and preliminary evaluation of multi-scale searching. *In CHI '07 : Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1215–1224, New York, NY, USA. ACM. (Cité page 74)

Plaisant, C., Carr, D. et Shneiderman, B. (1995). Image browsers : Taxonomy, guidelines, and informal specifications. Rapport technique. (Cité page 12)

Potamianos, G., Lamel, L., Wolfel, M., Huang, J., Marcheret, E., Barras, C., McDonough, J., Hernando, J., Macho, D. et Nadeu, C. (2009). *Computers in the Human Interaction Loop*. Chapitre 6. Automatic Speech Recognition., pages 43–59. Springer. (Cité page 157)

Prinzmetal, W., McCool, C. et Park, S. (2005). Attention : reaction time and accuracy reveal different mechanisms. *J Exp Psychol Gen*, 134(1):73–92. (Cité page 147)

Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466. (Cité page 44)

Rébillat, M. (2011). *Vibrations de plaques multi-excitateurs de grandes dimensions pour la création d'environnements virtuels audio-visuels. Approches acoustique, mécanique et perceptive.* Thèse de doctorat, Ecole Polytechnique. (Cité pages 42 et 51)

Riede, T., Herzel, H., Hammerschmidt, K., Brunnberg, L. et Tembrock, G. (2001). The harmonic-to-noise ratio applied to dog barks. *Journal of the Acoustic Society of America*, 110(4). (Cité page 93)

Robinson, C. W. et Sloutsky, V. M. (2004). The effect of stimulus familiarity on modality dominance. *In 26th Annual Meeting of the Cognitive Science Society (CogSci2004)*. (Cité page 57)

Robinson, C. W. et Sloutsky, V. M. (2010). Effects of multimodal presentation and stimulus familiarity on auditory and visual processing. *Journal of Experimental Child Psychology*, 107:351–358. (Cité page 57)

Rosenbaum, R. et Schumann, H. (2009). Resource-saving image browsing based on jpeg2000, blurring, and progression. *In Multimedia on Mobile Devices 2009*. (Cité pages 17, 29, 34, et 123)

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C. et Foxe, J. J. (2007). Do you see what i'm saying ? exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, 17:1147–1153. (Cité pages 49, 95, et 99)

Sakoe, H. et Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49. (Cité page 135)

Sarkar, M. et Brown, M. H. (1994). Graphical fisheye views. *Commun. ACM*, 37(12):73–83. (Cité pages 15, 34, et 66)

Schmandt, C. (1998). Audio hallway : a virtual acoustic environment for browsing. *In ACM UIST Symposium on User Interface Software and Technology*, pages 163–170. ACM Press. (Cité page 21)

Schmandt, C. et Mullins, A. (1995). Audiostreamer : Exploiting simultaneity for listening. *In Conference companion on Human factors in computing systems*, CHI '95, pages 218–219, New York, NY, USA. ACM. (Cité pages 21 et 24)

Schneider, T. R., Engel, A. K. et Debener, S. (2008). Multisensory identification of natural objects in a two way crossmodal priming paradigm. *Experimental Psychology*, 55(2):121–132. (Cité pages 50, 98, 100, 107, 116, 117, 118, et 126)

Schrammel, J., Giller, V., Tscheligi, M., Kosara, R., Hauser, H. et Miksch, S. (2003). Experimental evaluation of semantic depth of field, a preattentive method for focus+context visualization. *In Human-Computer Interaction – INTERACT'03*, pages 888–891. (Cité pages 17, 123, 126, et 130)

Schröger, E. et Widmann, A. (1998). Speeded responses to audiovisual signal changes result from bimodal integration. *Psychophysiology*, 35:755–759. (Cité page 51)

Sekuler, R., Sekuler, A. B. et Lau, R. (1997). Sound alters visual motion perception. *Nature*, 385(308). (Cité page 49)

Serrano, M. (2010). *Interaction Multimodale en Entrée : Conception et Prototypage*. Thèse de doctorat, Université de Grenoble. (Cité page 165)

Shafiro, V. (2008a). Development of a large-item environmental sound test and the effects of short-term training with spectrally-degraded stimuli. *Ear and Hearing*, 29(5):775–790. (Cité pages 91 et 95)

Shafiro, V. (2008b). Identification of environmental sounds with varying spectral resolution. *Ear and Hearing*, 29(3):401–420. (Cité pages 91, 93, 95, et 124)

Shams, L., Kamitani, Y. et Shimojo, . (2002). Visual illusion induced by sound. *Cognitive Brain Research*, 14. (Cité page 49)

Shams, L. et Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews*, 7(3):269–284. (Cité page 47)

Shen, J. et Reingold, E. M. (2001). Visual search asymmetry : The influence of stimulus familiarity and low-level features. *Perception and Psychophysics*, 63:463–475. (Cité page 33)

Shinn-Cunningham, B. (2002). Speech intelligibility, spatial unmasking, and realism in reverberant spatial auditory displays. *In International Conference on Auditory Display (ICAD'2002)*. (Cité page 124)

Simpson, B., Iyer, N. et Brungart, D. S. (2010). Aurally aided visual search with multiple audio cues. *In International Conference on Auditory Displays (ICAD'10)*, pages 51–54. (Cité pages 34 et 52)

Soto, R., López, M., Diego, D. D. et Manuel, G. (2008). Absolute threshold of coherence of position perception between auditory and visual sources for dialog. *In 125th Convention of the AES*. (Cité page 54)

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology*, 28(2):61–70. (Cité pages 47 et 56)

Spence, C. (2011). Crossmodal correspondences : A tutorial review. *Attention, Perception, & Psychophysics*, 73:971–995. 10.3758/s13414-010-0073-7. (Cité page 47)

Spence, C. et Driver, J. (1994). Covert spatial orienting in audition : exogeneous and endogeneous mechanisms. *Journal of Experiemental Psychology : Human Perception and Performances*, 20:555–574. (Cité page 38)

Spence, C. et Driver, J. (1997). On measuring selective attention to an expected sensory modality. *Perception & Psychophysic*, 59:389 – 403. (Cité page 57)

Spence, C. et Driver, J. (2000). Attracting attention to the illusory location of a sound : reflexive crossmodal orienting and ventriloquism. *Neuroreport*, 11:2057–2061. (Cité page 49)

Spence, R. (2001). *Information Visualization*. Chapitre 7. Presentation, pages 111–133. ACM Press. (Cité page 12)

Spence, R. (2002). Rapid, serial and visual presentation : a presentation technique with potential. *Information Visualization*, 1:13–19. (Cité page 17)

Spieth, W., Curtis, J. F. et Webster, J. C. (1954). Responding to one of two simultaneous messages. *Journal of the Acoustic Society of America*, 26:391–396. (Cité page 145)

Stein, B. E. et Meredith, M. A. (1993). *The merging of the senses*. MIT Press, Cambridge (MA), US. (Cité pages 47, 54, 106, et 155)

Suh, B., Woodruff, A., Rosenholtz, R. et Glass, A. (2002). Popout prism : Adding perceptual principles to overview+detail document intsrfaces. *In CHI 2002*, Minneapolis, Minnesota, USA. (Cité page 34)

Suied, C., Bonneel, N. et Viaud-Delmon, I. (2007). Role of semantic vs spatial congruency in a bimodal go/no-go task. Poster. (Cité pages 55 et 116)

Suied, C., Bonneel, N. et Viaud-Delmon, I. (2009). Integration of auditory and visual information in the recognition of realistic objects. *Experimental Brain Research*, 194:91–102. (Cité pages 50, 98, 100, et 116)

Suied, C., Susini, P., McAdams, S. et Patterson, R. D. (2010). Why are natural sounds detected faster than pips ? *The Journal of the Acoustical Society of America*, 127. (Cité page 118)

Suied, C. et Viaud-Delmon, I. (2009). Auditory-visual object recognition time suggests specific processing for animal sounds. *PLoS ONE*, 4(4). (Cité page 117)

Sumby, W. et Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26:212–215. (Cité pages 49, 95, et 106)

Theeuwes, J. (2004). Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review*, 11(1):65–70. (Cité page 141)

Thorpe, S., Fize, D. et Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381:520–522. (Cité page 100)

Treisman, A. (1986). Features and objects in visual processing. *Scientific American*, 255(5): 114B–125. (Cité page 30)

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London : Series B, Biological Science*, 353:1295–1306. (Cité page 53)

Treisman, A. et Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12:97–136. (Cité page 30)

Treisman, A. et Gormican, S. (1988). Feature analysis in early vision : Evidence from search asymmetries. *Psychological Review*, 95:14–48. (Cité page 33)

Van der Burg, E., Olivers, C. N. et Bronkhorst, A. (2008). Pip and pop : Non spatial auditory signals improve spatial visual search. *Journal of Experimental Psychology : Human Perception and Performance*, 34(5):1053–1065. (Cité pages 34 et 52)

Varghese, L. A., Ozmeral, E. J., Best, V. et Shinn-Cunningham, B. G. (2012). How visual cues for when to listen aid selective auditory attention. *Journal of the Association for Research in Otolaryngology (JARO)*, online. (Cité page 53)

Vatakis, A. et Spence, C. (2008). Evaluating the influence of the "unity assumption" on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica*, 127(1):12–23. (Cité page 56)

Walker, B. N., Nance, A. et Lindsay, J. (2006). Spearcons : speech-based earcons improve navigation performance in auditory menus. *In Proceedings of the 12th International Conference on Auditory Display*, pages 63–68, London, UK. Department of Computer Science, Queen Mary, University of London, UK. (Cité page 24)

Wang, Q., Cavanagh, P. et Green, M. (1994). Familiarity and pop-out in visual search. *Perception and Psychophysics*, 56:495–500. (Cité page 33)

Ware, C. (2000). *Information visualization : perception for design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. (Cité pages 30, 31, 33, et 34)

Ware, C. et Lewis, M. (1995). The dragmag image magnifier. *In Conference companion on Human factors in computing systems*, CHI '95, pages 407–408, New York, NY, USA. ACM. (Cité page 14)

Watanabe, K. (2001). *Crossmodal Interaction in Humans*. Thèse de doctorat, California Institute of Technology, Pasadena, Californie, USA. (Cité page 49)

Welch, R. et Warren, D. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88:638–667. (Cité pages 55 et 56)

Wolfe, J. M. (2001). Asymmetries in visual search : An introduction. *Perception and Psychophysics*, 63(3):381–389. (Cité page 33)

Wood, N. et Cowan, N. (1995). The cocktail party phenomenon revisited : how frequent are attention shifts to one's name in an irrelevant auditory channel ? *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 21:255–260. (Cité page 39)

Woods, D.L. andAlain, C., Diaz, R. et Rhodes, D. andOgawa, K. (2001). Location and frequency cues in auditory selective attention. *J Exp Psychol Hum Percept Perform*, 27(1):65–74. (Cité page 38)

Wright, M., Freed, A. et Momeni, A. (2003). Opensound control : State of the art 2003. *In 2003 Conference on NIME*, pages 153–159, Montreal, Canada. (Cité page 73)

Yamaashi, K., Tani, M. et Tanikoshi, K. (1993). Fisheye videos : distorting multiple videos in space and time domain according to users' interests. *In CHI '93 : INTERACT '93 and CHI '93 conference companion on Human factors in computing systems*, pages 119–120, New York, NY, USA. ACM. (Cité page 26)

Yamamoto, D., Ozeki, S. et Takahashi, N. (2009). Focus+glue+context : an improved fisheye approach for web map services. *In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 101–110, New York, NY, USA. ACM. (Cité page 66)

Yumoto, E., Gould, W. J. et Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustic Society of America*, 71:1544–1550. (Cité page 93)

Yuval-Greenberg, S. et Deouell, L. Y. (2009). The dog's meow : asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research*, 193(4):603–614. (Cité pages 55 et 56)

Zahorik, P., Brungart, D. et Bronkhorst, A. W. (2005). Auditory distance perception in humans : a summary of past and present research. *Acta Acustica united with Acustica*, 91(3):409–420. (Cité page 42)

Zanella, A., Rounding, M. et Carpendale, M. S. T. (2000). On the effects of visual cues in comprehending distortions. Rapport technique, University of Calgary, Department of Computing Science. (Cité pages 34 et 68)

Zhao, J., Chevalier, F., Pietriga, E. et Balakrishnan, R. (2011). Exploratory analysis of time-series with chronolenses. *IEEE Transactions on Visualization and Computer Graphics*, 17:2422–2431. (Cité page 12)

Zölzer, U., Amatriain, X., Arfib, D., Bonada, J., De Poli, G., Dutilleux, P., Evangelista, G., Keiler, F., Loscos, A., Rocchesso, D., Sandler, M., Serra, X. et Todoroff, T. (2002). *DAFX : Digital Audio Effects*. John Wiley & Sons. (Cité page 124)